O'REILLY®
Technical Guide

Data Sharing and Collaboration with Delta Sharing

Accelerate Innovation and Generate New Revenue Streams

Compliments of
databricks

Ron L'Esteve

databricks DELTA SHARING

The open, seamless and secure way to share data

Databricks launched Delta Sharing — the industry's first open protocol for secure data sharing across platforms, clouds and regions.

Learn how 6,000+ organizations have adopted Delta Sharing to accelerate innovation.

Get started

Today's economy revolves around data

Data sharing is essential to drive business value for data-driven organizations, but Gartner predicts only 15% will succeed.*

Develop a plan for success by:

- Adopting an open and secure approach to data sharing
- Putting your data to work for you
- Discovering insights faster

*Source: Gartner

Data Sharing and Collaboration with Delta Sharing

Accelerate Innovation and Generate New Revenue Streams

Ron L'Esteve

Beijing · Boston · Farnham · Sebastopol · Tokyo

O'REILLY®

Data Sharing and Collaboration with Delta Sharing
by Ron L'Esteve

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisition Editor: Aaron Black
Development Editor: Gary O'Brien
Production Editor: Kristen Brown
Copyeditor: Arthur Johnson

Proofreader: Helena Stirling
Interior Designer: David Futato
Cover Designer: Ellie Volckhausen
Illustrator: Kate Dullea

March 2024: First Edition

Revision History for the First Edition
2024-03-25: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781098160227* for release details.

The views expressed in this work are those of the author and do not represent the publisher's views. While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

This work is part of a collaboration between O'Reilly and Databricks. See our statement of editorial independence.

978-1-098-16023-4

[LSI]

Table of Contents

CHAPTER 1

Harnessing the Power of Data Sharing

Within the dynamic landscape of modern business, data has emerged as the cornerstone of innovation and growth. The pursuit of knowledge has never been more paramount, and organizations are navigating an ever-expanding ocean of data. Data sharing is the practice of making data available to other individuals or organizations for various purposes, such as analysis, research, innovation, or decision making. Data sharing can enable collaboration across different domains and sectors, creating new opportunities and insights. The advent of data sharing has brought forth an opportunity that transcends traditional boundaries, presenting a transformative path to propel businesses into new frontiers. Organizations recognize the immense potential of data sharing, as it enables them to gain valuable insights, make informed decisions, and drive innovation. Collaborative efforts among companies, researchers, and institutions have led to groundbreaking advancements and solutions to complex challenges in healthcare, finance, and education, among other domains.

Data sharing in the context of an organization can be defined both internally and externally. *Internal data sharing* refers to the sharing of data between different departments or teams within the organization itself. For example, the sales department might share customer data with the marketing department to help it create targeted campaigns. This type of sharing is crucial for collaboration and can lead to more informed decision-making processes. On the other hand, *external data sharing* involves making data available to entities outside of the organization, including partners, customers, researchers, or even the public. For instance, a healthcare organization might share anonymized patient data with research institutions to aid in medical research. With both types of sharing, it’s important to have clear policies and procedures in place to protect sensitive information and comply with relevant data protection regulations.

This chapter will explore the power of data sharing and how it can overcome various challenges to unlock its full potential. The chapter begins by examining the current landscape of data sharing, in which different types of data are generated and stored by various actors. You will learn about the main challenges hindering data sharing, such as legacy solutions that are not designed for interoperability and scalability, cloud vendors that create silos and lock-in effects, and legal and regulatory barriers that limit data access and reuse. You will also learn about use cases that demonstrate the value and impact of data sharing in different domains, such as health, education, agriculture, energy, transportation, and societal good. The chapter will also highlight some of the principles and best practices that enable effective and responsible data sharing and collaboration.

Finally, you will explore the benefits of data sharing and collaboration for various stakeholders by demonstrating how data sharing can improve efficiency, innovation, transparency, accountability, trust, and participation. The risks and challenges of data sharing in relation to privacy, security, quality, fairness, and sovereignty will also be addressed. The chapter will conclude by providing some recommendations and guidelines for fostering a culture of data sharing and collaboration that balances the benefits and risks of data sharing.

Data sharing is a key capability for digital transformation, and data and analytics leaders who share data internally and externally are likely to generate measurable economic benefit. The market for data sharing is also growing rapidly as organizations realize the need for an ecosystem that enables them to build, sell, and buy data products. Forbes estimates that by 2030, $3.6 trillion dollars will be generated through the commercialization of data products. According to Gartner, chief data officers (CDOs) who have successfully executed data sharing initiatives in their organizations are 1.7 times more effective at showing business value and return on investment from their data analytics strategy. Gartner predicts that by 2024, most organizations will attempt trust-based data sharing programs, but only 15% will succeed and outperform their peers on most business metrics. Gartner also predicts that by 2026, 15% of large enterprises will have evaluated connected governance to effectively manage complex cross-organizational challenges with governance programs.[1]

Current Data Sharing Models

The landscape of data sharing is constantly evolving and is presently marked by a surge in the volume of data being generated, collected, and stored by organizations across various industries. The advent of advanced technologies, including cloud computing and the Internet of Things (IoT), has fueled this data explosion, creating a wealth of information waiting to be harnessed. New technologies and platforms are emerging to facilitate data sharing between organizations while preserving privacy and security. According to Boston Consulting Group, the models most likely to facilitate broad data sharing within and across industries are vertical platforms, super platforms, shared infrastructure, and decentralized models. These models offer different approaches to centralizing or decentralizing data, with varying benefits and risks. Let's briefly go over the four models:

Vertical platforms

Vertical platforms provide data sharing and solutions to targeted needs within individual industries, such as predictive maintenance, supply chain optimization, operational efficiency, and network optimization. Examples of vertical platforms include Airbus's Skywise and Penske's Fleet Insight. Skywise allows airlines, maintenance providers, and other aviation stakeholders to share data and collaborate on solutions for predictive maintenance, supply chain optimization, and operational efficiency. Fleet Insight platform provides data sharing and analytics solutions for the transportation and logistics industry, allowing fleet operators to share data and collaborate on solutions for network optimization, route planning, and asset utilization.

[1] Andrew White, "Our Top Data and Analytics Predictions for 2021," *Gartner Business Insights, Strategies & Trends for Executives* (blog), January 12, 2021.

Super platforms

Super platforms centralize data from multiple sources and domains and provide services and applications based on the aggregated data. These platforms provide a wide range of services, such as data storage, processing, analytics, machine learning, and artificial intelligence, that allow organizations to share and analyze their data in a centralized manner. Examples of super platforms include major cloud providers such as Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform, Alibaba Cloud, IBM Cloud, Tencent Cloud, OVHcloud, DigitalOcean, and Akamai Connected Cloud.

Shared infrastructure

Shared infrastructure centralizes data from multiple sources and domains but differs from super platforms in that it potentially raises fewer issues regarding control over the market, access to the data, and the value the data generates. An example of shared infrastructure is the European Union's Gaia-X initiative, which aims to create a federated cloud infrastructure for Europe that allows organizations to share data securely and transparently while retaining control over their data. The initiative is designed to promote data sharing and collaboration while addressing concerns about market dominance by large technology companies.

Decentralized models

In decentralized models, data resides at its source and is accessible to others without central intermediaries. Decentralized models use blockchain, peer-to-peer networks, or distributed ledgers to enable secure and transparent data sharing. Data mesh is an example of a decentralized model that organizes data by a specific business domain—for example, marketing, sales, customer service, and more—providing more ownership to the producers of a given dataset. Data mesh is an architectural pattern for implementing enterprise data platforms in large and complex organizations and helps scale analytics adoption beyond a single platform and implementation team. Ocean Protocol and IOTA are other examples of decentralized models that use blockchain and distributed ledger technologies respectively to enable secure data sharing between organizations and devices.

Deloitte Insights highlights the rise of powerful data sharing and privacy-preserving technologies that are ushering in a new era of data monetization.[2] Advances in data sharing technologies have enabled the buying and selling of potentially valuable information assets in highly efficient, cloud-based marketplaces. Privacy-preserving technologies such as fully homomorphic encryption (FHE) and differential privacy allow for the sharing of encrypted data and for computations to be performed without decrypting the data first. This has fueled a promising trend in which stores of sensitive data that were previously unused due to privacy or regulatory concerns are now generating value across enterprises in the form of new business models and opportunities.

Data Exchanges and Marketplaces

Emerging technologies and platforms allow for secure data sharing while preserving privacy and security. These capabilities allow organizations to share data with other individuals or organizations for analysis, research, innovation, decision making, and more. Some of these platforms include data marketplaces and data exchanges. A *data marketplace* is an online platform where users can buy or sell different types of datasets and data streams from several sources. Data marketplaces are mostly cloud services through which individuals or businesses upload data to the cloud. Data marketplaces can be useful for organizations that want to monetize their data or access external data sources to enhance their analytics and decision-making capabilities. A *data exchange* is a platform that facilitates data sharing among various stakeholders. Data exchanges provide access to data points from around the world to fuel data-driven marketing activities and advertising, which can be useful for organizations that want to collaborate with other organizations in their industry or across industries to create new insights and opportunities. Table 1-1 shows a few examples of cloud platforms that offer data sharing and marketplace capabilities.

[2] "Data-Sharing Technologies Made Easy," Deloitte Insights, December 7, 2021, *https://oreil.ly/-EbDl*.

Table 1-1. Cloud platforms with data sharing and marketplaces

Platform	Capabilities
AWS Data Exchange	Facilitates discovery and utilization of third-party data in the cloud. Supports various data types, including datafiles, tables, and APIs. Streamlines data procurement and governance.
Azure Data Share	Provides secure sharing of data snapshots from Azure SQL Database and Azure Synapse Analytics to other Azure subscriptions. Supports sharing comprehensive data snapshots from various SQL resources in Azure. Enables quick, direct access to data.
Databricks Delta Sharing and Marketplace	Delta Sharing offers an open protocol for secure real-time data sharing with other organizations across platforms, clouds, and regions; this allows sharing of table collections in a Unity Catalog metastore without the need for copying. Unity Catalog simplifies access management, enhances security with fine-grained control, and harnesses AI to automate monitoring and uphold data and ML model quality. Databricks Marketplace is powered by Delta Sharing and offers datasets, AI models, and prebuilt notebooks.
Snowflake Direct Share and Marketplace	Facilitates secure direct data sharing between Snowflake accounts with read-only access to all shared database objects. The Snowflake Marketplace connects providers and consumers, offering ready-to-use data, services, and Snowflake Native Apps.

Secured Platforms

Secured platforms are technologies that provide a safe and secure environment for organizations to share data with partners or customers while maintaining control over their information. Some examples of secured platforms include marketplaces and data exchange platforms offered by major cloud providers, as well as clean rooms, blockchain, and distributed ledgers. These technologies offer benefits such as increased security, transparency, efficiency, and cost savings. In this section, you will learn about the following technologies and how they can help organizations to securely share data and collaborate with partners:

Clean rooms

A data clean room is a secure collaboration environment that allows two or more participants to leverage data assets for specific, mutually agreed upon uses, while guaranteeing enforcement of strict data access limitations—e.g., not revealing or exposing their customers' personal data to other parties.[3]

[3] "Data Clean Rooms: Guidance and Recommended Practices," IAB Tech Lab, accessed October 27, 2023, *https://oreil.ly/yhrG2*.

Companies use this technology to share data with partners or customers securely and efficiently. Clean rooms have many use cases, including compliance with privacy laws, data anonymization, profile enrichment, customer overlap observation, and score analysis. A sample use case for clean rooms could include *federated learning*, a machine learning technique that allows multiple organizations to collaboratively train a machine learning model without sharing their raw training data. This can enable organizations to benefit from each other's expertise while still protecting the privacy of their individual datasets. Amazon Marketing Cloud, Databricks Clean Rooms, Disney Clean Room Solution, Google Ads Data Hub, Habu, LiveRamp Safe Haven, and Snowflake are some of the major cloud providers that offer clean room services.

Blockchain

Blockchain is a type of distributed ledger technology that uses cryptography to create a secure and transparent record of transactions. A distributed ledger is a digital system in which records of transactions are simultaneously maintained at multiple points throughout a network, eliminating the need for a central authority to keep a check against manipulation. Blockchain allows multiple parties to share data without the need for a central authority. Blockchain has many use cases, including supply chain monitoring, payment processing and money transfers, digital identity management, royalty protection and copyright management, cybersecurity, digital voting, real estate management, and healthcare data management. Some major cloud providers that offer blockchain services are Amazon Managed Blockchain, Bloq, Corda, Dragonchain, IBM Blockchain Platform, Kaleido, and Oracle Blockchain Cloud Services.

Distributed ledger technology

Distributed ledger technology (DLT) is a digital database technology that records transactions in a distributed ledger of computer networks to enhance the secure, decentralized, and transparent transmission of data. DLT has many use cases, including recording transactions, securing identities, collecting votes, entering contracts, and demonstrating ownership. Major cloud providers that offer distributed ledger services include Microsoft Azure confidential ledger.

Industry Use Cases

Data sharing can enable collaboration across different domains and sectors, creating new opportunities and insights that can benefit society, the economy, and the environment. Data sharing can also improve efficiency, innovation, transparency, accountability, trust, and participation. Here are some use cases of data sharing:

Healthcare and life sciences

By enabling medical researchers worldwide to collaborate in real time, data sharing can accelerate breakthroughs in areas such as COVID-19 vaccine development, ultimately saving lives. Moreover, data sharing empowers healthcare providers, payers, researchers, and patients to harness insights from diverse sources such as electronic health records, clinical trials, wearable devices, and genomic data, leading to improved patient outcomes, reduced costs, and enhanced quality of care.

Financial services

In the financial sector, data sharing brings a new level of efficiency and security. Financial institutions can strengthen fraud detection and regulatory compliance by sharing insights and transactions securely with regulators and auditors, contributing to economic stability. Similarly, farmers can make informed decisions by accessing weather, soil, and market data, optimizing their crops and yields through smart farming.

Education

Within education, data sharing enhances student learning and engagement. Educators, students, parents, and administrators can utilize academic records, assessments, and attendance and behavior data to create tailored learning experiences.

Transportation

The transportation sector benefits from data sharing through smart mobility solutions. By leveraging data from traffic, weather, location, routes, fares, and preferences, transportation providers, users, regulators, and planners can enhance travel experiences. This optimization contributes to reduced congestion, emissions, accidents, and costs, which results in smoother and more efficient transportation systems.

Energy

Data sharing underpins the evolution of energy management. Smart grids become possible when energy producers, consumers, distributors, and regulators access generation, consumption, distribution, and pricing data. This facilitates the integration of renewable energy sources, curbing emissions, boosting energy efficiency, and ensuring grid reliability.

Retail/consumer goods

Retail experiences are transformed with data sharing, offering personalized shopping journeys. Retailers can tailor interactions based on purchase history, preferences, and behavior, which enriches customer experiences. Data sharing streamlines inventory management, supply chain coordination, and pricing strategies when the data is shared with suppliers, distributors, and partners.

Manufacturing

The manufacturing industry embraces data sharing for smarter production processes. Manufacturers leverage data from sensors, machines, and processes to enhance product quality, reduce costs, and foster innovation. Sharing data with suppliers, customers, and partners leads to collaborative innovation and more efficient production cycles.

Government/public sector

Data sharing elevates governance in the digital age. Government agencies can harness data from citizens, businesses, and other agencies to improve public services and transparency. Through shared data, governments can enhance transparency, reduce corruption, and make more informed decisions to better serve citizens and stakeholders**.**

Communications, media, and entertainment

Data sharing reshapes media consumption with personalized experiences. Media companies leverage data on viewing history, preferences, and behavior to curate content tailored to individual tastes. Beyond personalization, data sharing enhances content creation, distribution, and monetization strategies, enabling media companies to connect more effectively with creators, advertisers, and partners.

Data sharing can create value by enabling new insights, solutions, and opportunities. In the next section, you will learn about the benefits of data sharing and collaboration.

Benefits of Data Sharing and Collaboration

Sharing data is a crucial aspect of digital transformation that can provide significant advantages to a wide range of organizations, both internally and externally. As per Gartner, leaders in data and analytics who share data externally reap three times more tangible economic benefits than those who don't.[4] Data sharing can enhance the relevance of data, leading to the generation of more substantial data and analytics to address business issues and achieve organizational objectives. Encouraging data sharing can lead to greater engagement and influence among stakeholders. By making data available to other individuals or organizations for various purposes, such as analysis, research, innovation, or decision making, data sharing can enable collaboration across different domains and sectors, creating new opportunities for businesses. One of the key benefits of data sharing and collaboration is the ability to create new insights. By pooling data from multiple sources, organizations can discover new patterns, trends, and relationships in data that can lead to novel insights, solutions, and opportunities.

Data sharing can eliminate inefficiencies, resulting in improved financial health, enhanced collaboration, and the creation of new opportunities among business leaders. It can also improve policy formulation and foster trust within organizations. Data sharing enables organizations to access valuable data assets, boost collaboration and productivity among employees, cut down unnecessary efforts and expenses, enhance transparency among stakeholders, and facilitate more efficient resource utilization.

Data sharing and collaboration revolutionize the way organizations operate, unlocking new insights and driving innovation. Collaborative data sharing can improve efficiency by reducing duplication, redundancy, and waste. Organizations can optimize their resources, operations, and services by allowing data-driven decision making through access to diverse and rich data sources. Data sharing

[4] Lawrence Goasduff, "Data Sharing Is a Business Necessity to Accelerate Digital Business," Gartner, May 20, 2021, *https://oreil.ly/olfUQ*.

and collaboration also encourage innovation. Organizations can foster creativity, learning, and experimentation by allowing access to diverse perspectives and rich data sources. This can lead to the development of new products, services, or business models that create value for customers, stakeholders, and society. For example, in retail, personalized shopping experiences can be enabled by allowing retailers to access purchase history, preferences, and behavior data.

Allowing access to data provenance, lineage, metadata, or documentation can increase trust and transparency. This improves data collection, processing, and analysis visibility and accountability. It also improves the quality and reliability of data by allowing verification, validation, and feedback. For instance, fraud detection and safeguarding economies can be enhanced by allowing financial institutions to securely share insights and transactions with regulators, auditors, and other parties. Trust between data providers and consumers can be built by establishing clear roles, responsibilities, rights, and obligations. Meeting ethical, legal, social, and technical standards can also ensure trust within the data ecosystem.

Finally, data sharing and collaboration can empower participation by enabling data access and use for education, research, innovation, advocacy, or other purposes. It also enables data contribution and feedback from various actors such as citizens, businesses, academia, civil society, and individuals. The US Department of Education has developed a toolkit to help communities leverage relationships while protecting student privacy.[5] This toolkit is designed to inform civic and community leaders who wish to use shared data to improve academic and life outcomes for students. Within the research domain, data sharing enabled more than three thousand Chinese companies to rapidly scale the production of medical supplies during the COVID-19 pandemic.[6] The World Economic Forum discussed how secure data sharing can enable innovation and stated that 76% of business executives agree that innovation requires new ways

[5] US Department of Education, *Data-Sharing Tool Kit for Communities: How to Leverage Community Relationships While Protecting Student Privacy* (Washington, DC: 2016), *https://oreil.ly/P4EWv*.

[6] Shameen Prashantham and Jonathan Woetzel, "3 Lessons from Chinese Firms on Effective Digital Collaboration," *Harvard Business Review*, August 10, 2020, *https://oreil.ly/BNGs6*.

of collaborating with ecosystem partners and third-party organizations.[7] The importance of data sharing in advocacy efforts is highlighted by a report published by the Harvard Kennedy School on how civil society and business are engaging in joint advocacy to change policy, attitudes, and practices.[8]

However, it's important to note that while data sharing and collaboration can bring about significant benefits, they also have ethical, legal, social, and technical implications. These include issues related to obtaining informed consent for data sharing, anonymizing data, controlling access to data, working with research ethics committees and institutional review boards, and more. Therefore, careful consideration is needed when sharing data with others.

In summary, data sharing offers numerous benefits to a wide array of organizations, both internally and externally. These benefits include unlocking valuable data assets, fostering collaboration and productivity, reducing inefficiencies, enhancing transparency and trust, and promoting efficient resource utilization. Data sharing is an essential business requirement for accelerating digital business transformation.

Challenges

Factors that hinder data sharing include legacy solutions that are not designed for interoperability and scalability, cloud vendors that create silos and lock-in effects, legal and regulatory barriers that limit data access and reuse, lack of standardization in data formats, issues related to data ownership and control, and the costs and resources required to share data. Companies may be deterred from taking advantage of data sharing opportunities by challenges and risks in such areas as the following:

[7] Hugo Ceulemans, Mathieu Galtier, Tinne Boeckx, Marion Oberhuber, and Victor Dillard, "From Competition to Collaboration: How Secure Data Sharing Can Enable Innovation," World Economic Forum, June 27, 2021, *https://oreil.ly/boydV*.

[8] Richard Gilbert and Jane Nelson, *Advocating Together for the SDGs—How Civil Society and Business Are Joining Voices to Change Policy, Attitudes and Practices* (Cambridge, MA: Business Fights Poverty and the Corporate Responsibility Initiative at the Harvard Kennedy School, 2018), *https://oreil.ly/FhsJf*.

Regulations

One of the challenges of data sharing is complying with the various regulations designed to protect the rights and interests of individuals and organizations. Data sharing is subject to different laws and rules depending on the type, source, and destination of the data and the purpose and context of the data sharing. (See the sidebar "Key Regulations Governing Data Sharing in Different Regions and Domains" on page 16 for examples of key regulations; please note that it is not an exhaustive list, so be sure to do your due diligence about local and industry-specific regulations as required.)

Privacy

Another challenge of data sharing is keeping personal or sensitive data safe from unauthorized access or use. Sharing data may involve disclosing personal or sensitive information to third parties, which can raise ethical and legal concerns. For instance, sharing data may violate the consent or preferences of the data subjects, expose them to identity theft or discrimination, or harm their reputation or dignity. Sharing data may also conflict with existing regulations or standards, such as the General Data Protection Regulation (GDPR) in the European Union, which imposes strict rules on how personal data can be collected, processed, and transferred.

Security

Data must be protected from unauthorized modification, deletion, or disclosure. Sharing data may involve transferring data over networks or storing data in cloud platforms, which can expose data to cyberattacks or breaches. For example, sharing data may compromise data confidentiality, integrity, or availability, resulting in data loss, corruption, or leakage. Sharing data may also require implementing appropriate encryption, authentication, authorization, or auditing mechanisms to ensure data security.

Quality

Data quality refers to the accuracy, completeness, consistency, timeliness, and relevance of data. Sharing data may involve integrating data from multiple sources or domains, which can introduce quality issues. For example, sharing data may lead to errors in the data itself such as inconsistencies, duplicates, missing values, or outdated values. Sharing data may also require certifying the origin of the data and its documentation to ensure its quality.

Trust

Trust refers to the confidence and reliability of data and its sources. Data sharing may involve collaborating with unknown or untrusted parties, which may affect trust. For example, data sharing may face challenges in verifying the identity, reputation, or credibility of data providers or consumers. Data sharing may also require establishing mechanisms such as contracts, agreements, incentives, ratings, reviews, or feedback to ensure trust.

The current challenges of data sharing require careful consideration and management by data owners and users. Data owners need to balance the benefits and risks of sharing their data with others while respecting the rights and interests of the data subjects. Data users need to assess the quality and trustworthiness of the shared data while complying with the terms and conditions of the data providers. Data sharing also requires adopting appropriate technologies and platforms that facilitate secure and efficient data sharing while preserving privacy and quality.

It is important to ensure that data sharing is done responsibly and securely. This can be achieved through the use of data access controls, data encryption, data masking, data classification, data governance policies, and data sharing agreements. Such practices help ensure that shared data is protected and used ethically and transparently. Let's review these practices:

Data access controls

Mechanisms that restrict access to sensitive data to only authorized users, which can help prevent unauthorized access and misuse of the data.

Data encryption

The process of converting plain text into a coded format that can be read only by someone with the key to decrypt it, thus protecting the confidentiality of the data while it is being transmitted or stored.

Data masking

The process of obscuring sensitive information in a dataset so it cannot be easily identified, protecting the privacy of individuals whose information is included in the dataset.

Data classification
: The process of organizing data into categories based on its level of sensitivity, ensuring that appropriate security measures are applied to different types of data.

Data governance policies
: Rules and procedures for managing and protecting shared data, which can help ensure that the data is used responsibly and ethically.

Data sharing agreements
: Legal contracts that specify the terms and conditions under which data will be shared between parties, ensuring that all parties understand their rights and responsibilities concerning the shared data.

Despite the challenges of sharing data efficiently, it is possible to maintain privacy and security while still taking advantage of opportunities for innovation through responsible data stewardship. As organizations continue to navigate the landscape of data sharing, finding the right balance between open collaboration and responsible stewardship is key to maximizing its benefits.

Summary

In this chapter, you have learned about the power of data sharing and how it can transform the way you create, access, use, and share data. You also learned about the benefits and challenges of data sharing and collaboration for various stakeholders and society. You examined the current landscape of data sharing, in which various actors, such as governments, businesses, academic organizations, civil society, and individuals, generate and store different types of data. You also looked at the main models that facilitate broad data sharing within and across industries, such as vertical platforms, super platforms, shared infrastructure, and decentralized models.

However, data sharing also requires adopting appropriate technologies and platforms that facilitate secure and efficient data sharing while preserving privacy and quality. Databricks provides the first open source approach to data sharing and collaboration across data, analytics, and AI with products that include Delta Sharing, Unity Catalog, Marketplace, and Clean Rooms. Delta Sharing provides an open protocol for secure cross-platform live data sharing. It integrates with Unity Catalog for centralized governance to manage access control policies. This allows organizations to share live data across platforms without replicating it. With its Marketplace, users can discover, evaluate, and access data products—including datasets, machine learning models, dashboards, and notebooks—from anywhere without the need to be on the Databricks platform. Clean Rooms provides a secure environment for businesses to collaborate with their customers and partners on any cloud in a privacy-safe way. Participants in the data clean rooms can share and join their existing data and run complex workloads in any language—Python, R, SQL, Java, Scala—on the data while maintaining data privacy.

In Chapter 2, you will dive deeper into Delta Sharing and how it works. You will learn about the features and benefits of Delta Sharing, such as open cross-platform sharing, avoiding vendor lock-in, and easily sharing existing data in Delta Lake and Apache Parquet formats to any data platform. You will also learn how to use Delta Sharing to share and consume data from various sources and destinations, such as Databricks, AWS S3, Azure Blob Storage, Google Cloud Storage, Snowflake, Redshift, BigQuery, Presto, Trino, and Spark SQL. In addition, you will explore some of the use cases and best practices of Delta Sharing in different domains and scenarios. By the end of Chapter 2, you will have a solid understanding of Delta Sharing and how it can enable effective and responsible data sharing and collaboration across different platforms and domains.

Key Regulations Governing Data Sharing in Different Regions and Domains

The Data Act

The EU's Data Act is a legislative proposal adopted by the European Commission in December 2021 that aims to create a common European data space in which nonpersonal data can flow freely across borders and sectors. *Nonpersonal data* refers to data that is not related to an identified or identifiable natural person, such as industrial data, public sector data, or environmental data.

The General Data Protection Regulation (GDPR)

The GDPR is a comprehensive regulation that governs the processing of personal data within the EU. *Personal data* refers to any information relating to an identified or identifiable natural person, such as name, email address, location, or biometric data. The GDPR sets strict rules for how personal data can be collected, used, and shared, and it gives individuals greater control over their personal information.

The California Consumer Privacy Act (CCPA)
The CCPA is a comprehensive data privacy law enacted in California in 2018. It gives California residents the right to know what personal information is being collected about them, the right to request that their personal information be deleted, and the right to opt out of the sale of their personal information.

The Health Insurance Portability and Accountability Act (HIPAA)
HIPAA is a federal law enacted in the United States in 1996. It sets national standards for protecting the privacy and security of individuals' health information. HIPAA applies to covered entities, such as healthcare providers, health plans, and healthcare clearinghouses, as well as their business associates.

CHAPTER 2
Understanding Delta Sharing

Delta Sharing is an innovation from Databricks that revolutionizes how organizations share and exchange data. It provides a simple, secure, and open way for data providers and consumers to share data across organizations in real time, regardless of which computing platforms they use. This protocol is built on top of Delta Lake, an open source storage layer that brings reliability to data lakes. It provides ACID (atomicity, consistency, isolation, durability) transactions and scalable metadata handling and unifies streaming and batch data processing on top of an existing data lake. Unlike traditional methods, Delta Sharing employs a streamlined REST (representational state transfer) protocol to grant access to specific segments of cloud datasets. REST serves as the cornerstone of web service creation, enabling data transmission over HTTP in a lightweight manner. Delta Sharing natively integrates with cloud storage systems such as Amazon S3, Azure Data Lake Storage (ADLS), Cloudflare R2, and Google Cloud Storage (GCS), ensuring seamless and dependable data transfer between data providers and recipients.

Delta Sharing fundamentally simplifies how data is accessed. Whether working with live data, notebooks, dashboards, or sophisticated models like machine learning and AI, Delta Sharing enables secure sharing from your lakehouse to any computing environment. Notably, it frees data from the confines of proprietary systems, enabling data sharing in versatile formats such as Delta Lake and Apache Parquet. Delta Sharing has the unique capability to share live data effortlessly across various data platforms, cloud environments, or geographical regions—without the need for replication.

Delta Sharing sets itself apart from other solutions through its dedication to an open exchange of data. “Open” in the context of Delta Sharing refers primarily to the open source nature of its sharing protocol, which promotes a wide network of connectors and ensures superior interoperability across diverse platforms. While Delta Lake does employ an open datafile format, the key differentiator is the open source sharing protocol that enables this expansive network. The term “open exchange” is used here to denote a marketplace that isn’t limited to a single vendor, in contrast to a “private exchange.” Despite the democratization of data access, Delta Sharing maintains stringent security, governance, and auditing mechanisms. Capable of managing massive datasets, Delta Sharing scales seamlessly, marking a significant advancement in data sharing and accessibility.

In this chapter, you will learn about the features and capabilities of Delta Sharing, how Delta Sharing fits into the broader Databricks ecosystem, the advantages of using Delta Sharing over traditional data sharing methods, and real-world use cases in which Delta Sharing can be applied. By the end of the chapter, you will have a solid understanding of what Databricks Delta Sharing is and how to get started with data sharing. In addition, you’ll discover how strategic partnerships with popular enterprise companies such as Oracle and Cloudflare enhance Delta Sharing. Let’s explore the capabilities of Databricks Delta Sharing, a technology that facilitates effective collaboration and sets the stage for a business model that thrives on shared data and enhanced possibilities. In gaining an understanding of its capabilities, you’ll learn how Delta Sharing can augment data sharing and teamwork within your organization.

Features of Delta Sharing

The Delta Sharing ecosystem includes a wide range of components that work together to enable secure data sharing across platforms, clouds, and regions:

Cloud storage systems

Utilizing modern cloud storage systems like S3, ADLS, Cloudflare R2, and GCS, Delta Sharing enables efficient and reliable distribution of large-scale datasets.

Support for diverse clients

Data recipients can directly interface with Delta Shares using various popular libraries and programming languages, including pandas, Apache Spark, and Rust, enhancing the accessibility of shared data and removing the necessity for a specific computational setup.

Security and governance

Delta Sharing provides robust security and governance features that allow data providers to govern, track, and audit access to shared datasets easily, ensuring that shared data is accessed only by authorized recipients and that usage is tracked for compliance purposes.

Scalability

Delta Sharing is designed to share terabyte-scale datasets efficiently and reliably by using cloud storage systems like S3, Cloudflare R2, ADLS, and GCS to transfer data between data providers and recipients.

Open source project

Delta Sharing also exists as an open source initiative, enabling the sharing of Delta tables across various platforms.

Integration with Unity Catalog

The inherent compatibility of Delta Sharing with Unity Catalog provides a unified platform for managing, overseeing, auditing, and monitoring the usage of shared data.

Shares and recipients

The main concepts behind Delta Sharing in Databricks are shares and recipients. A *share* is a group of data assets that are read-only and can be shared with one or more recipients. A *recipient* is an entity that links an organization with a credential or a secure sharing identifier that grants that organization access to one or more shares.

Delta Sharing's Technical Capabilities

Delta Sharing is available in two versions: an open source version and a managed version. The open source version is a reference sharing server that you can use to share Delta tables from other platforms; the managed version is provided by Unity Catalog, which allows you to provide different sets of users fine-grained access to any datasets from one central place.

The key distinction between the two versions is in the degree of management and control they offer. In the open source version, you're tasked with segregating datasets with varying access rights across multiple sharing servers and enforcing access restrictions on those servers and their associated storage accounts. On the other hand, the managed version simplifies governance, tracking, and auditing of access to your shared datasets through Unity Catalog.

When choosing between the open source and managed versions of Delta Sharing, it is important to assess your requirements. If you need fine-grained control over data access and want to manage and audit data sharing easily, the managed version may be a better choice. The open source version may be sufficient if you are comfortable managing your own sharing servers and storage accounts.

Delta Sharing has three main components: providers, shares, and recipients. *Providers* are entities that have made data available for sharing. *Shares* are logical groupings of data assets that may include tables, notebook, volumes, ML models, or other assets that providers intend to share. *Recipients* are organizations that have been granted access to one or more shares. Delta Sharing can be utilized in two distinct manners: open sharing and Databricks-to-Databricks sharing. Open sharing provides the ability to share data with any user, regardless of their access to Databricks, by using a token-based credential. On the other hand, Databricks-to-Databricks sharing allows for data sharing with Databricks users who are connected to a different metastore than yours, utilizing a secure sharing identifier. Databricks-to-Databricks sharing also supports notebook sharing, AI models, and view sharing. Delta Sharing supports Spark Structured Streaming and Delta Lake time travel queries. It also has some limitations, such as only supporting tables in Delta format, having a reserved schema name (`information_schema`), and having quotas for Delta Sharing resources.

In this section, you will learn about several additional components you can integrate with Delta Sharing, including Unity Catalog, Databricks Clean Rooms, Structured Streaming, view and notebook sharing, and more.

Unity Catalog

Delta Sharing can share collections of data assets stored in a Unity Catalog metastore in real time without copying them. Unity Catalog is a metastore service provided by Databricks that stores metadata about Delta tables, such as their schema, partitioning information, and access control lists. This capability is available for both internal data sharing within your organization and Databricks-to-Databricks data sharing with partners that have their own Databricks workspace. Unity Catalog can also be used on the provider side for Databricks-to-open sharing, in which case the provider can still get the governance benefit of Unity Catalog, but the recipient won't. By sharing access to the tables in the Unity Catalog, Delta Sharing allows recipients to immediately begin working with the latest version of the shared data without waiting for the data to be copied or replicated.

Unity Catalogs understand all types of data and the computations done with them in the Databricks platform, providing a uniform way to do access control, lineage, discovery, monitoring, data sharing, and more. Anup Segu, co-head of data engineering at YipitData, published a blog post on the Databricks website in April 2023 describing how YipitData, a data collaboration platform that analyzes over 15 petabytes of alternative data in its lakehouse, switched from using a Hive metastore to using Databricks Unity Catalog as a metastore service to scale its data operations and provide robust data governance controls. The article explains the challenges and limitations of the previous architecture, the benefits and features of Unity Catalog, and the outcomes and impacts of the migration. YipitData experienced the advantages of Unity Catalog, which simplified and reduced the cost of managing its data platform, enhanced the visibility and use of its data assets, improved the security and efficiency of data sharing, and created new ways to deliver data to its clients.

Recently, Databricks announced that Unity Catalog is adding an Apache Hive API, allowing any engine that understands Hive to talk to it and thus expanding Unity Catalog to have enterprise-wide reach. To use Delta Sharing, you need to enable it for your Unity Catalog metastore, create shares and recipients, grant access, and monitor data sharing. Recipients of the shared data can utilize a variety of tools and platforms for access, including but not limited to Apache Spark, pandas, Power BI, and Databricks. Figure 2-1 illustrates this end-to-end process flow.

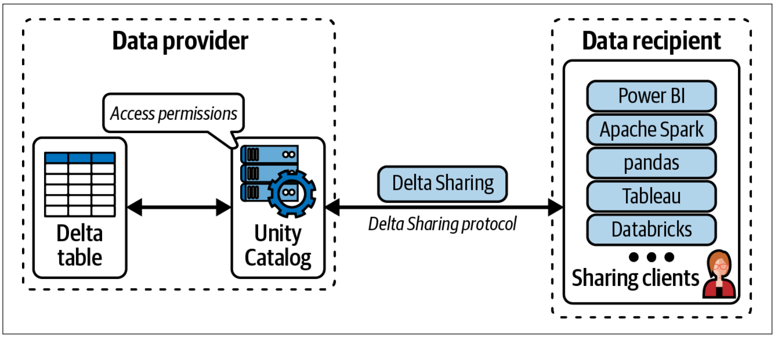


Figure 2-1. Delta Sharing architecture

Databricks Clean Rooms

A data clean room is a secure and isolated environment in which two or more parties can combine and process their respective datasets to produce a net new dataset. Databricks Clean Rooms is a clean room solution powered by Delta Sharing that enables secure and privacy-safe data collaboration across organizational boundaries. The term "privacy-safe" in this context refers to the protection of sensitive or personal information during the data collaboration process. The data clean room ensures that data is safeguarded and encrypted and is accessible only to authorized parties, and it allows control over the level of detail shared and the insights derived from the combined data. With Databricks Clean Rooms, you can share and join your data with multiple other organizations without exposing or replicating your data. You can run complex analytics and machine learning on the shared data using any language, such as SQL, R, Scala, Java, or Python. Databricks Clean Rooms excels in interoperability, ensuring smooth collaboration across diverse environments. With Delta Sharing, collaborators can collaborate across different cloud providers, regions, and even data platforms without requiring extensive data movement, eliminating data silos and enabling organizations to leverage existing infrastructure and data ecosystems while maintaining the utmost security and compliance.

Structured Streaming

Structured Streaming with Delta Sharing is now generally available on Azure, AWS, and GCP, allowing for practical real-time data integration. Data recipients on the Databricks Data Intelligence Platform can efficiently stream changes from Delta tables shared through the Unity Catalog. Data providers can leverage Structured Streaming to simplify data-as-a-service scaling, reduce sharing costs, perform immediate data validation, and enhance customer service with real-time data delivery. Meanwhile, data recipients can stay updated on shared datasets without costly batch processing while streamlining data access and enabling real-time applications.

Views and Notebook File Sharing

Databricks has added new collaboration features on top of the Delta Sharing protocol that allow you to share data and AI assets with customers and partners across organizations.

You can share any logical views of filtered and curated data using Delta Sharing without any extra data replication. Common use cases include data curation and abstraction, access control and data obfuscation, and data monetization for when you have tables with a large number of recipients. Using view sharing will only charge the recipient for the cost of the view computation. You can also share notebooks that allow you to combine data and code together so that data recipients can get insights quickly. Using notebook sharing, recipients can view the shared data alongside descriptions and examples. Your data recipients can view the notebooks or convert into their own notebooks.

In the Databricks Catalog Explorer, you can add both notebook files and views to a share. As the owner of the share with read permission on the notebook, you can add a notebook file, optionally assigning an alias that recipients will see and use. Any updates or deletions to a shared notebook file require re-adding it with a new alias or removing it from the share, potentially necessitating notification of the recipient.

Similarly, views, which are virtual tables derived from SQL queries, can be added to a share by selecting the desired share, clicking on the Assets tab, and then choosing Add view. You can then select the schema and view for sharing and optionally provide an alias for recipient access. The added view will be listed in the Views list on the Assets tab. Note that views can also be added using the Databricks Unity Catalog CLI or the `add view SQL` command.

AI Model Sharing

Delta Sharing's technical capabilities extend beyond data sharing to include AI model sharing. This feature is particularly beneficial for organizations that want to leverage machine learning models across different platforms and teams.

In the context of Delta Sharing, AI models are considered as data assets and can be included in shares, similar to tables, notebooks, and volumes. This allows providers to share AI models with recipients, facilitating collaboration and knowledge transfer.

The process of sharing AI models is similar to that for sharing other data assets. In the open source version of Delta Sharing, you would need to manage your own sharing servers and storage accounts to share AI models. However, the managed version simplifies this process by providing fine-grained access control and easy management through Unity Catalog.

Databricks-to-Databricks sharing supports AI model sharing, allowing you to share models with Databricks users connected to a different metastore. This is done using a secure sharing identifier, ensuring that your models are shared securely and only with authorized users.

Delta Sharing's AI-model-sharing capability enhances the platform's versatility, making it a powerful tool for organizations that want to democratize access to AI models while maintaining robust security and governance.

Time Travel

Delta Lake time travel is a powerful feature that enables you to access previous versions of a Delta table based on timestamps or specific table versions. Its functionality serves practical purposes such as recreating analyses, reports, or outputs, as well as facilitating auditing and data validation tasks. In the context of Delta Sharing, consider a scenario in which you have a shared Delta table containing sales data. You're interested in analyzing how sales figures evolved over the past year. With Delta Lake time travel, you can easily query the data as it existed at various points in time. For instance, you could retrieve the state of the data as of January 1, February 1, and so on, giving you the ability to track changes over time by accessing historical snapshots of the data.

Furthermore, if an unexpected issue arises, you can use time travel to roll back to a previous version of the data for recovery. It's worth noting that the retention of these historical versions is controlled by configuration settings, so it's essential to adjust those settings to ensure the availability of the desired data versions for querying your data.

Schema Evolution

Schema evolution allows you to evolve the schema of a Delta Lake table over time. You can add new columns to your table, change the data type of existing columns, or even delete columns without having to rewrite your entire dataset.

You can enable schema evolution by setting the `mergeSchema` option to `true` when writing data to a Delta table, allowing you to append data with a different schema to an existing Delta table. For example, if you have a Delta table with the columns `first_name` and `age` and want to append data that also includes a country column, you can do so by setting the `mergeSchema` option to `true`. The new column will be added to the Delta table, and any existing data in the table will have null values for the new column.

You can also enable schema evolution by default by setting the `autoMerge` option to `true`. This allows you to append data with different schemas without having to set the `mergeSchema` option every time.

Schema evolution allows you to easily adapt your data sharing practices as your business requirements change. Using schema evolution, you can ensure that your shared data remains relevant and up to date, without going through complex data migration processes.

Partition Filtering

In Delta Sharing, partition filtering allows data providers to share specific parts of a Delta table with data recipients without making extra copies, helping you share only the data you need or control access based on recipient characteristics. To specify a partition that filters by recipient properties when creating or updating a share, you can use Data Explorer, or the CURRENT_RECIPIENT SQL function in a Databricks notebook or in the Databricks SQL query editor. Default properties include databricks.accountId, which is the Databricks account that a data recipient belongs to (for Databricks-to-Databricks sharing only). Partition filtering enables cost-effective and secure high-quality data sharing at scale. It is especially useful when sharing data with multiple parties from a single table. Figure 2-2 illustrates how you can share data based on a single multitenant table with multiple recipients.

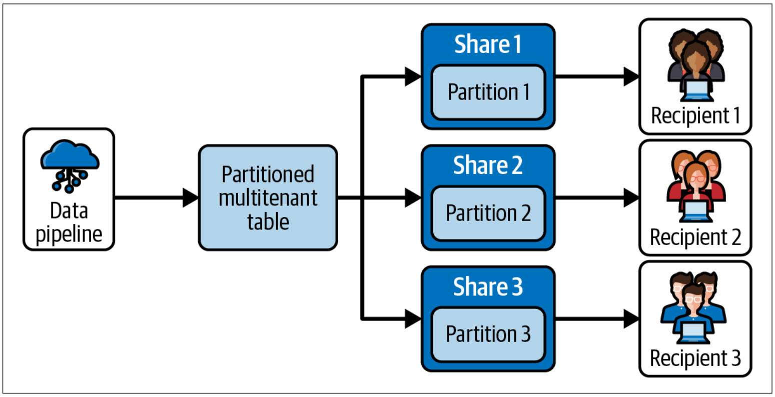


Figure 2-2. Partitioned multitenant static table sharing

Partition filtering can be further enhanced by using parameterized partition sharing, which allows data providers to reference the CURRENT_RECIPIENT function in the partition specification. This function is dynamically evaluated to a value of a data recipient's property and can be used to match partition columns. CURRENT_RECIPIENT supports two kinds of properties: *built-in properties* and *custom properties*. Built-in properties are predefined for every recipient object and cannot be modified. Custom properties are user-managed properties that the user can create, update, and delete. Figure 2-3 shows how you can dynamically share partitions with recipients using the CURRENT_RECIPIENT property value.

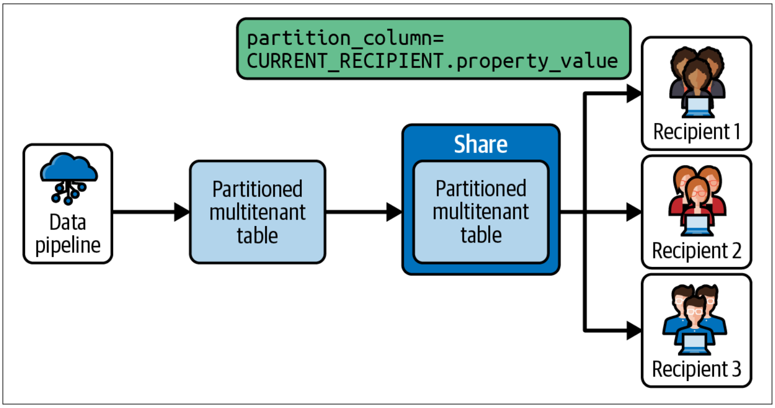

partition_column=
CURRENT_RECIPIENT.property_value
Data pipeline
Partitioned multitenant table
Share
Partitioned multitenant table
Recipient 1
Recipient 2
Recipient 3

Figure 2-3. Partitioned multitenant dynamic table sharing

Advantages of Secure Cross-Platform Data Sharing

An open protocol for secure cross-platform data sharing such as Delta Sharing provides many advantages. It allows organizations to share data with other organizations easily, regardless of which computing platforms they use, simplifying the data sharing process and eliminating the need for data duplication.

Secure data sharing is a multifaceted process. It begins with data providers having the ability to control access to their data, with the power to grant and revoke access as needed. This ensures that data is only in the hands of those who are authorized to view it.

In addition, data providers can share specific partitions of data. This means that only relevant and necessary data is shared with recipients, enhancing both data privacy and data quality.

Last, secure data sharing involves the implementation of dynamic and flexible access control rules. These rules are based on the recipient's attributes, such as their name, account ID, or partner ID. This level of customization allows for a more secure and efficient data sharing process. Secure cross-platform data sharing enhances data collaboration and productivity by allowing organizations to govern, track, and audit access to their shared datasets while collaborating with their customers and partners on any cloud in a privacy-safe environment. Dynamic access control also enables organizations to securely share data from their data lakes without data replication.

Collaborators can meet on their preferred cloud and have the flexibility to run complex computations and workloads in any language, such as SQL, R, Scala, Java, or Python.

In addition, secure cross-platform data sharing provides strong security, governance, and auditing while scaling to handle massive datasets. Sharing data in its original format, without the need for data duplication or transformation, preserves the integrity and accuracy of the data, along with reducing the risk of errors or inconsistencies.

Comparison with Other Data Sharing Solutions

Unlike traditional data sharing solutions that are often tied to a single vendor or cloud provider, Delta Sharing lets you avoid vendor lock-in and share data across different data platforms, clouds, or regions without replicating or copying it to another system. You can share data with your customers and partners in a privacy-safe environment while maintaining full control and governance over your shared datasets.

Delta Sharing also provides several unique features and benefits that make data sharing faster, easier, and more secure—for example:

- Data recipients can directly connect to Delta Shares from familiar tools and open source frameworks without having to first deploy a specific compute pattern, making it easy to access and analyze shared data using familiar tools and frameworks.
- Data providers can share existing data in Delta Lake and Apache Parquet formats without the need to transform or export the data to another format, preserving the formats' schema, metadata, and performance benefits while enabling cross-platform compatibility and also reducing the complexity of ELT, manual sharing, and lock-ins to a single platform.
- Data sharing is based on an open protocol that supports standard authentication methods such as OAuth 2.0, which ensures that only authorized users can access the shared data and that the data is encrypted both in transit and at rest.

- Data sharing is integrated with Databricks Unity Catalog, a global data catalog that allows you to discover, browse, and request access to shared datasets from various sources. You can also use the Unity Catalog to manage your own shares and track their usage and performance.

Delta Sharing allows you to unlock the full potential of your data lake and collaborate with your customers and partners on live data without compromising on security, performance, or flexibility.

Key Partnership Integrations

To expand its data sharing ecosystem and reach more data providers and consumers, Databricks has partnered with technology providers such as Cloudflare, Dell, Oracle, and Twilio. These partnerships aim to enhance Delta Sharing's capabilities, provide mutual benefits for both Databricks and its partners, and enable consumers to access shared data from multiple platforms without being constrained to a specific vendor. Let's look more closely at the technology providers just mentioned:

Cloudflare R2

Cloudflare R2 offers zero egress fees or provider compute as well as a distributed object storage solution, and it seamlessly integrates with Databricks Delta Sharing. Using Cloudflare R2's capabilities, data providers can share lakehouse data without egress fees, making it a cost-effective solution. Its integration with Databricks Delta Sharing enables data providers to share lakehouse data rapidly, reducing costs associated with data transfer and enhancing user satisfaction through efficient data accessibility.

Dell Technologies Cloud Storage for Multi-Cloud

Dell Technologies Cloud Storage for Multi-Cloud is a fully managed service that uses Delta Lake to orchestrate data analytics across multiple cloud domains. Data providers can extend their lakehouse data stored within Dell Technologies Cloud Storage for Multi-Cloud to data consumers. The protocol allows data consumers to interface with shared data through a number of systems, or to directly query it using Spark SQL. The integration enables organizations to conduct multicloud analytics, leading to cost savings from efficient data management and improved user satisfaction due to the speed of data sharing.

Oracle Autonomous Database

Oracle Autonomous Database is a self-driving, self-securing, and self-repairing cloud database service. It plays a dual role within the Delta Sharing landscape, as both a data provider and a data recipient. Oracle Autonomous Database enables data sharing from its lakehouses by securely exchanging data with data consumers through the Delta Sharing protocol. Compatibility with the protocol provides data consumers with access to shared data, either through Delta Sharing–friendly systems or by directly querying it using Oracle SQL. This allows secure and swift data exchange with data consumers, providing them with quick access to shared data. This speed of data sharing enhances user satisfaction and can lead to cost savings by reducing the time spent on data retrieval.

Twilio Segment

Twilio Segment is a customer data platform that helps businesses collect, clean, and activate their customer data. It has partnered with Databricks to enable businesses to use customer data from Twilio Segment in the Databricks lakehouse for AI applications. The partnership offers a bidirectional integration that allows sending and activating data between Twilio Segment and Databricks, as well as creating golden customer profiles with Profiles Sync. The benefits of the partnership include improved customer experience, personalized marketing campaigns, real-time communication, upsell and churn reduction, cost and efficiency optimization, and more. The partnership leverages Twilio Segment's real-time CDP and Databricks's advanced machine learning capabilities, enhancing user satisfaction by delivering actionable insights and personalized experiences at scale.

These partnerships not only enhance Delta Sharing's capabilities but also provide significant benefits to end users in terms of speed of data sharing, cost savings, and user satisfaction. They enable more data sources and destinations, enhance performance and reliability, and support diverse use cases.

Use Cases: Real-World Applications of Data Sharing

Delta Sharing is a powerful and flexible data sharing solution that enables organizations to exchange data with customers, suppliers, and partners securely. It is used in various industries to improve collaboration, streamline operations, and drive innovation. For example, a large retailer used Delta Sharing to easily share product data with more than one hundred partners across different cloud platforms, without having to replicate the data across regions. Similarly, a manufacturer used Delta Sharing to govern and share data across distinct internal entities, without having to move data, thus improving collaboration and compliance with data regulations. Healthcare providers use Delta Sharing to share patient data with researchers and collaborators securely and efficiently while complying with HIPAA and other regulations, enabling faster and more accurate diagnosis and treatment of diseases. Media companies can use Delta Sharing to share content and analytics data with their advertisers and partners, without compromising on data quality or security, resulting in increased revenue and customer loyalty. These are just some of the real-world examples of how Delta Sharing has been used to solve data sharing challenges. Here are some additional examples of how Delta Sharing can be used:

Data commercialization

Financial data providers can utilize Delta Sharing to share large datasets seamlessly and overcome scalability issues with their SFTP servers, resulting in improved customer satisfaction and reduced operational costs.

Data sharing with external partners and customers

Delta Sharing can be used to share data with external partners and customers securely. For example, a company can share sales data with its suppliers to improve supply chain efficiency, or it can share customer data with its partners to develop better marketing campaigns. Delta Sharing provides strong security, governance, and auditing while scaling to handle massive datasets, ensuring that sensitive data is protected and shared in compliance with regulations.

Line of business data sharing

Delta Sharing can be used for internal line of business sharing. For example, a company can use Delta Sharing to securely share data with business units and subsidiaries across clouds or regions. This enables seamless collaboration among different business units and improves decision making and insights.

These are just a few examples of how Delta Sharing can be used in the real world. Its flexibility, scalability, and security make it an ideal solution for organizations looking to improve collaboration and drive innovation through secure and real-time data sharing.

Data Governance and Security in Delta Sharing

Data governance and security are essential aspects of Delta Sharing, as they ensure that data is shared in a controlled and compliant manner. Delta Sharing allows you to share data securely and efficiently using the Data Intelligence Platform. It lets you control and monitor how your data is accessed by different users and organizations. Some of the key features that enable data governance and security in Delta Sharing are:

Fine-grained access control

You can specify who can access which tables and columns in your shares and apply row-level and column-level security policies. Your data consumers can use their own credentials to access your shares without your having to create or manage any accounts on your side.

Data encryption

Incorporating security from the ground up, this design ensures end-to-end TLS encryption, spanning from the client to the server and all the way to the storage account. This is similar to how you would secure your data both in transit and at rest using your own unique encryption keys, which remain confidential and are never shared with data consumers or third parties.

Short-lived credentials, such as presigned URLs, are used for data access. This is comparable to data consumers decrypting data using their own unique keys, which are also kept confidential. This strategy effectively shields your data from unauthorized access and tampering.

Additionally, you can easily govern, track, and audit access to your shared datasets. This feature gives you complete control over who can access your data and when, ensuring a secure and efficient data sharing process.

Data lineage

You can track the origin and history of your data, including who created, modified, or deleted it, and when and why. Your data consumers can also view the metadata and schema of your shared data, as well as any changes made to them over time. This ensures data quality and auditability.

Data compliance

You can comply with various data regulations and standards, such as the GDPR, the CCPA, HIPAA, the Payment Card Industry Data Security Standard (PCI DSS), and so on, by applying appropriate data policies and practices. Your data consumers can also follow their own compliance requirements by accessing only the data that they are authorized to use, resulting in reduced risk of data breaches and fines.

Delta Sharing Methods

Delta Sharing offers two main methods of data sharing—sharing between Databricks environments and sharing from Databricks to open source platforms—but overall there are four modes, including open-to-open and open-to-Databricks sharing.

Sharing Between Databricks Environments (Databricks to Databricks—D2D)

In this method, the recipient provides a unique identifier linked to their Databricks workspace. The data provider then creates a "share" in their own workspace, which includes tables, views, and notebooks. They also create a "recipient" object representing the user or group who will access the data. The provider grants access to the share, which then appears in the recipient's workspace. Users can access the share through various means, such as Catalog Explorer, the Databricks CLI, or SQL commands.

Sharing from Databricks to Open Source (Databricks to Open—D2O)

In this model, the data provider creates a “recipient” object and a “share” object, as in the previous method. However, in this case a token and an activation link are generated for the recipient. The provider sends the activation link to the recipient securely. The recipient uses this link to download a credential file, which is used to establish a secure connection with the provider and access the shared data. This method allows data to be read on any platform or tool.

Sharing from Open Source to Open Source (Open to Open—O2O)

This method allows data sharing between any open source platforms or tools, without requiring Databricks. The data provider can use an open source reference server to create and manage shares and recipients. The data recipient can use any Delta Sharing client to access the shared data using a credential file. This method enables data sharing across different clouds and regions, with minimal setup and maintenance.

Sharing from Open Source to Databricks (Open to Databricks—O2D)

Delta Sharing’s open-to-Databricks sharing feature enables data and AI model sharing beyond the Databricks ecosystem. This is achieved through a token-based credential system, allowing data providers to share assets with any user, irrespective of their Databricks access. For instance, when a customer shares data from Oracle to Databricks, it’s an example of open-to-Databricks sharing. Despite the openness, Delta Sharing ensures robust security and governance.

Key Differences

The key differences among the four methods of data sharing are based on the unique features of each method and the specific needs each caters to. Sharing between Databricks environments (D2D) is an excellent choice when you want to share data with another Databricks user, regardless of their account or cloud host. This method provides a seamless sharing experience within the Databricks ecosystem.

On the other hand, sharing from Databricks to open source (D2O) platforms offers more flexibility. It allows data sharing with any user on any platform, making it a more universal solution. This method is particularly useful when you want to share data with users who are using different open source tools or platforms.

Similarly, sharing from open source to Databricks (O2D) enables data sharing beyond the Databricks ecosystem. This method allows data providers to share data with any user, irrespective of their Databricks access, using a token-based credential system. This method is beneficial for data providers who want to leverage Databricks's advanced machine learning capabilities and share their AI models with data consumers.

Last, sharing from open source to open source (O2O) is the most flexible method. It allows data sharing between any open source platforms or tools, without requiring Databricks. This method enables data sharing across different clouds and regions, with minimal setup and maintenance. It's the go-to choice when both the data provider and the recipient are using open source tools.

Scenario: Sharing and Consuming Sales Data with Delta Sharing

Suppose you work for a company that wants to share its sales data with a partner company for analysis. You have the sales data stored in a Delta table in your Databricks workspace, and you want to use Delta Sharing to share it securely with the partner company. In this section, you will learn about the steps you can follow to share data using Delta Sharing. As a prerequisite, you'll need to make yourself a metastore admin or user with the `CREATE SHARE` privilege for the metastore by running the following commands:

```
-- Grant privileges to the role
GRANT CREATE SHARE ON metastore TO metastore_admin;
GRANT CREATE RECIPIENT ON metastore TO metastore_admin;

-- Grant the role to specific users
GRANT metastore_admin TO user1, user2;
```

Sharing Data

Figure 2-4 illustrates the five steps for sharing data with Delta Sharing. This section also breaks them out separately and includes code samples to help you learn how to create a new share, add Delta Lake tables to the share, create a recipient, share the activation link with customers, and define access levels, along with which data to share.

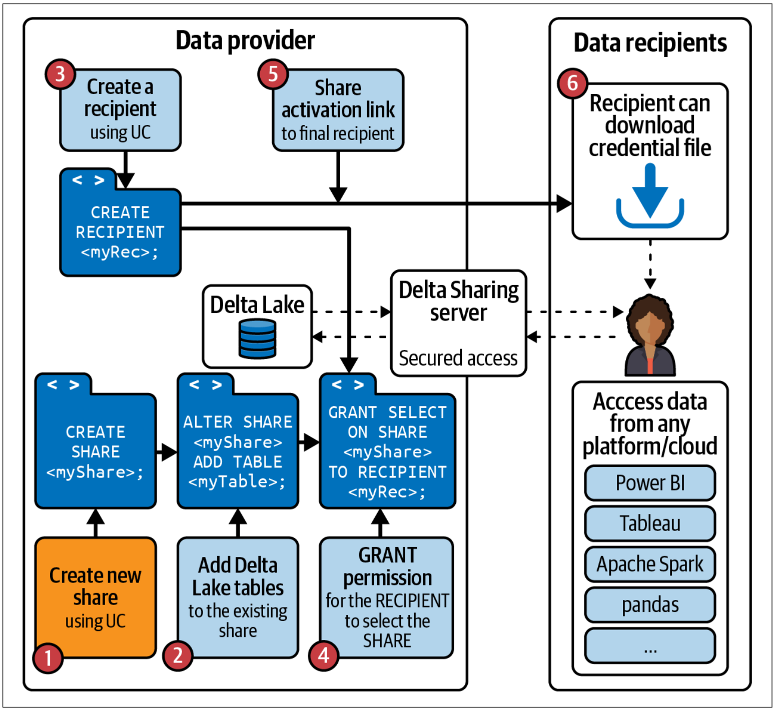

Data provider
3
Create a recipient using UC
5
Share activation link to final recipient
< >
CREATE RECIPIENT <myRec>;
Delta Lake
Delta Sharing server
Secured access
< >
CREATE SHARE <myShare>;
< >
ALTER SHARE <myShare> ADD TABLE <myTable>;
< >
GRANT SELECT ON SHARE <myShare> TO RECIPIENT <myRec>;
Create new share using UC
1
Add Delta Lake tables to the existing share
2
GRANT permission for the RECIPIENT to select the SHARE
4
Data recipients
6
Recipient can download credential file
Acccess data from any platform/cloud
Power BI
Tableau
Apache Spark
pandas
...

Figure 2-4. Delta Sharing process flow from data provider to recipient(s)

1. *Create a new share.*

   Run the following command in a Databricks notebook or the Databricks SQL query editor to create a new share:

```
CREATE SHARE IF NOT EXISTS US_Sales_Data
COMMENT 'Daily US Sales Data including all historical
sales’;
```

2. *Add Delta Lake tables to the share.*

   Next, run the following command to add tables to the share:

   ```
   ALTER TABLE tbl_USSalesData SET TBLPROPERTIES \
                (delta.enableChangeDataFeed = true);
   ALTER SHARE US_Sales_Data ADD TABLES tbl_USSalesData WITH
   CHANGE DATA FEED;
   PARTITION (SalesRegion = "SE", year LIKE "2000%") as
   tbl_USSalesData.`2000_sales`;
   ```

   Notice how you can set the table properties to enable change data feed and create the share with change data feed enabled. You can also select just a subset of the data by using the `PARTITION` function. Also, you are not limited to sharing tables; you can share views and notebooks as well.

3. *Create a recipient.*

   You can have multiple RECIPIENT(s) and assign them to multiple SHARE(s). Run the following command to create a recipient for your open sharing workflow:

   ```
   CREATE RECIPIENT IF NOT EXISTS US_Sales_Data_recipient
   COMMENT 'Recipient for my external customer using Open
   source Activation Link';
   ```

   Note that for sharing within Databricks environments, you'll also need to complete the additional step of having the recipient provide a unique sharing identifier for their Databricks workspace. This identifier is linked to the Unity Catalog metastore in the recipient's workspace. This metastore ID will not be relevant for the open sharing workflow. In the following code, you'll need to specify the unique metastore ID for your Databricks workspace recipient:

   ```
   CREATE RECIPIENT IF NOT EXISTS US_Sales_Data_recipient
   USING ID 'aws:us-west-2:<UUID>
   COMMENT 'Recipient for my external customer using Data-
   bricks';
   ```

4. *Grant access to the share.*

   Since you now have a RECIPIENT and a SHARE, you'll need to ensure your RECIPIENT has SELECT access to your SHARE. You can do this by running the following SQL command:

   ```
   GRANT SELECT ON SHARE US_Sales_Data TO RECIPIENT
   US_Sales_Data_recipient;
   ```

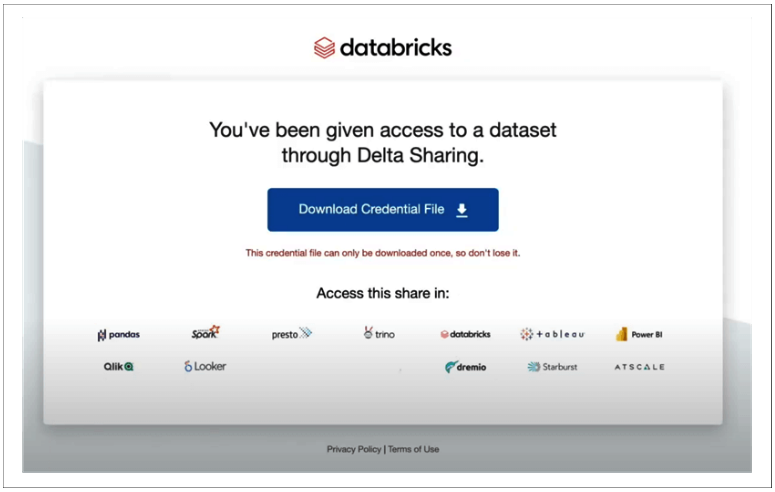
5. *Share the activation link with customers.*

   Each recipient of the open source sharing workflow is given an activation link that they can use to download their credentials. The file contains the Delta Server identification details and the authorized consumer. Note that the activation link is single use. This step is unique to the open source sharing workflow. In the D2D flow, the share appears automatically, simplifying the user experience and eliminating an extra step.

Consuming Data

To access shared data directly from Databricks (see step 6 in Figure 2-4), you need to have a Databricks workspace enabled for Unity Catalog and a unique sharing identifier for your workspace. Once you provide the identifier to the data provider, they will share data with you and create a secure connection with your organization. You can then find the share containing the data you want to access and create a catalog from it. After granting or requesting access to the catalog and its objects, you can read the data using various tools available to Databricks users. You can also preview and clone notebooks in the share.

Recipients who have received an activation link, similar to the illustration shown in Figure 2-5, can download the credential file locally in JSON format. Note that for security purposes, the credential file can be downloaded only one time, after which the download link is deactivated. For certain technologies, such as Tableau, in addition to the URL link, you may need to upload this credential file. For other technologies, you may need a “bearer token” or other credentials from this file.



You've been given access to a dataset through Delta Sharing.

This credential file can only be downloaded once, so don't lose it.

Access this share in:

Figure 2-5. Delta Sharing recipient activation link

Once the credential file has been downloaded, it can be utilized across various notebook platforms, such as Jupyter and Databricks, to access shared data stored within data frames. To enable this functionality in your notebook, execute the subsequent commands to install and import the Delta Sharing client. Alternatively, you may choose to install it from PyPi by searching for "delta-sharing-server" and installing the "delta-sharing" package. Following installation, you can employ the previously downloaded credential profile file to enumerate and access all shared tables within your notebook environment:

```
# Install the Delta Sharing Python Package
!pip install delta-sharing
# Import the Delta Sharing Libraries
import delta_sharing
# Define the Share File as a Dataframe that can be read by the
Delta
# Sharing Spark reader
spark_ussales_df = delta_sharing.load_as_spark("/FileStore/
    US_Sales_Data.share#US_Sales_Data")

# Analyze the shared sales data using Spark SQL
salesDF.createOrReplaceTempView("sales")
spark.sql("SELECT product, SUM(quantity)
          FROM sales
          GROUP BY product
          ORDER BY SUM(quantity) DESC").show()
```

Additionally, you have the option to utilize well-known reporting platforms like Power BI to access your shared tables. In the context of Power BI, connecting to a Delta Sharing source is a straightforward process. You can achieve this by selecting "Delta Sharing" from the readily available Power BI data source options and then clicking the Connect button, as demonstrated in Figure 2-6.

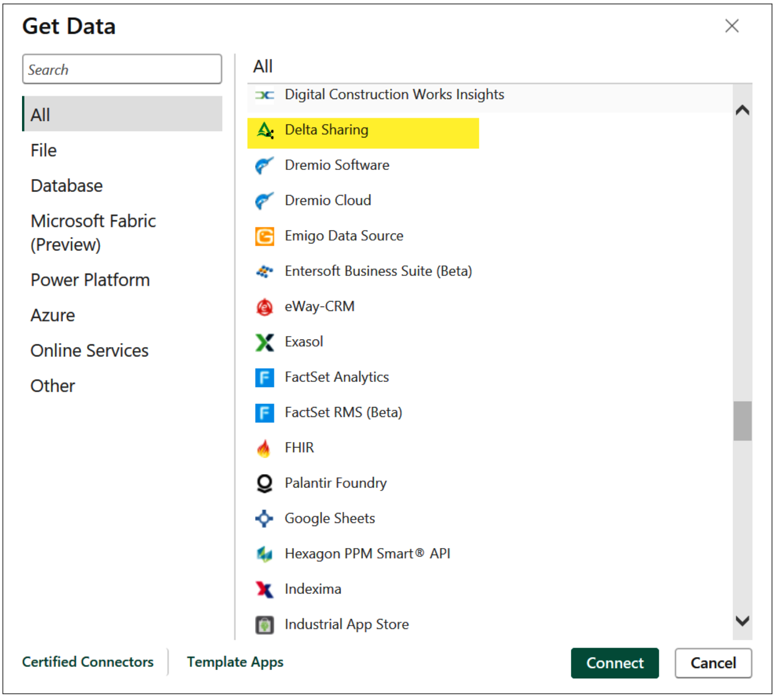


Figure 2-6. Connect to Delta Sharing with Power BI connector

In the configuration section, as depicted in Figure 2-7, you will need to enter the Delta Sharing server URL for authentication. Additionally, take note of the optional Advanced Options, which allow you to specify a row limit, enabling control over the number of rows retrieved from the source dataset.

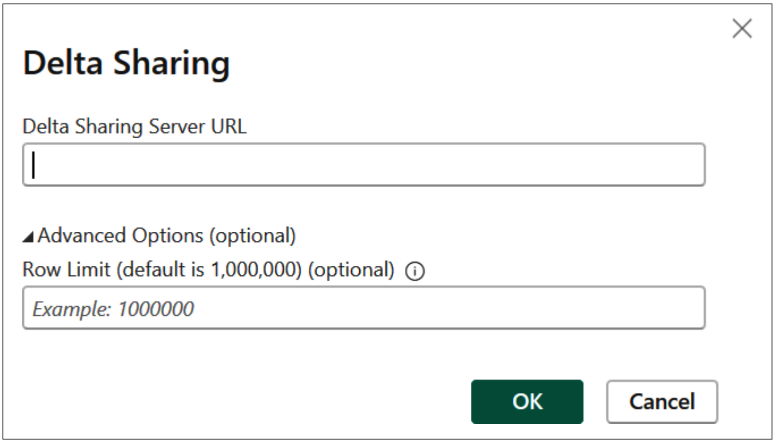

Delta Sharing
Delta Sharing Server URL
Advanced Options (optional)
Row Limit (default is 1,000,000) (optional)
Example: 1000000
OK
Cancel

Figure 2-7. Power BI connector configuration for Delta Sharing

Auditing Data Sharing

To audit data sharing using Delta Sharing, you can use Databricks audit logs. Here is an example of how to query the audit logs using SQL:

```
SELECT *
FROM audit_logs
WHERE actionName = '<action>'
```

This query will return all Delta Sharing events recorded in the audit logs. You can filter the audit logs by different actions, such as creating, modifying, deleting, or accessing shares and recipients. Remember to input the right actions within the `WHERE` clause placeholder. You can also view each event's request and response parameters, such as the `recipient name`, `metastore ID`, `share name`, `schema name`, `table name`, and `table version`. This information can help you identify errors and troubleshoot issues using the audit logs' status code and error message fields. For example, you can filter the results further by adding more conditions to the `WHERE` clause, such as `actionName = 'createShare'` to only return events related to creating shares. This command will return a list of all providers that you have access to. You can then use other CLI commands to view details about a specific provider, such as their name, description, and contact information.

In addition to filtering by action and viewing request and response parameters, you can use the `deltaSharingRecipientIdHash` field to correlate events across different recipients. This field is a hash of the `recipient ID` and can be used to identify all the events related to a specific recipient. You can also view the metadata of the share by running the following command: `DESCRIBE SHARE <share_name>;`. Finally, you can view all tables in a share by running the following command: `SHOW ALL IN SHARE <share_name>;`. By using these techniques, you can effectively audit data sharing using Delta Sharing and gain insights into how your data is being shared.

Best Practices

In addition to assessing the open source versus managed version based on your requirements, here are some best practices for securing your data sharing with Delta Sharing. By following these best practices, you can help ensure that your Delta Sharing platform is secure and that your data remains protected:

Set the appropriate recipient token lifetime for every metastore.
: To share data securely using the open sharing model, you need to manage tokens well. You need to set the default recipient token lifetime when you enable Delta Sharing for your Unity Catalog metastore if you want to use open sharing. Tokens should expire after a certain time period.

Establish a process for rotating credentials.
: It is important to establish a process for rotating the credentials (such as presigned URLs) used to access the data, which helps ensure that access to your data remains secure and that any compromised credentials are quickly invalidated.

Configure IP access lists.
: You can configure IP access lists to restrict access to your data based on the IP address of the client, ensuring that only authorized clients can access your data.

Enable audit logging.
: Enabling audit logging allows you to track and monitor access to your shared data while also identifying any unauthorized access or suspicious activity.

Summary

In this chapter, you have learned about the history and development of Delta Sharing, a data sharing solution that enables you to share live data directly from your data lake with anyone anywhere, on any platform. You have seen how Delta Sharing uses the open protocol for data exchange, which is based on the open source and open data formats of Delta Lake. You have also learned how Delta Sharing simplifies the data sharing process, enhances data collaboration and productivity, enables seamless data exchange, and strengthens data security and governance.

In reading about several use cases and advantages of Delta Sharing in various industries and scenarios, you have seen how Delta Sharing can be used for data commercialization, data sharing with external partners and customers, line of business data sharing, and more. You also have a deeper understanding of how Delta Sharing can help you avoid vendor lock-in, reduce operational costs, improve data quality, and enable cutting-edge data applications.

Key partnership integrations with Cloudflare, Dell, Oracle, and Twilio have enhanced Delta Sharing's capabilities and reach by enabling fluid data sharing between their platforms, Databricks, Apache Spark, pandas, Power BI, Excel, and other systems and frameworks that support the open protocol.

The data governance and security elements of Delta Sharing provided you with insights into its role in ensuring data privacy and compliance. By delving into the initial steps of getting started with Delta Sharing, you've learned about its capacity to streamline governance, tracking, and auditing of shared datasets. With Delta Sharing, you can collaborate securely with customers and partners across different cloud environments using Databricks Clean Rooms. You can define and manage access permissions, monitor data access, and track usage effectively.

Delta Sharing presents a transformative approach to data sharing and collaboration, offering simplicity, security, and scalability. It eliminates the need for data replication and facilitates sharing of existing data in Delta Lake and Apache Parquet formats across various platforms, clouds, and regions without tethering to a specific system. This flexibility extends to data recipients, who can seamlessly connect to Delta Shares from a range of systems and frameworks without requiring specialized compute patterns, marking a significant stride toward agile and efficient data sharing practices.

Delta Sharing is a powerful tool that can help you harness the power of data sharing and collaboration. Delta Sharing simplifies the data sharing process, enhances data collaboration and productivity, enables seamless data exchange, and strengthens data security and governance. You can also benefit from the open source and open data formats of Delta Lake to make data accessible to everyone.

In Chapter 3, you will delve into the practical aspects of navigating Databricks Marketplace. The Marketplace offers prebuilt integrations, data connectors, and AI models to enhance your data-driven initiatives. You will explore the various use cases and benefits it brings, enabling you to make informed decisions when utilizing its resources.

CHAPTER 3

Navigating the Open Data Marketplace

A data marketplace is a platform where data providers and data consumers come together to exchange data. It's essentially an online store for data, similar to how a traditional marketplace is a venue in which buyers and sellers meet to exchange goods. In a data marketplace, data providers are entities that have data to offer. These can be businesses, government organizations, research institutions, or even individuals who have collected and curated data. The data they provide can vary widely, from demographic data or industry-specific data to real-time sensor data and more. On the other hand, data consumers are entities that need data to drive their operations, research, or decision-making processes. They can be businesses looking for market trends, researchers in need of specific datasets, or developers building data-driven applications.

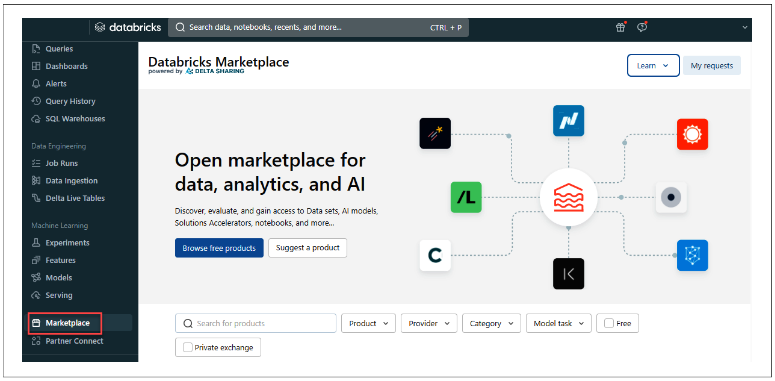
The data marketplace brings these two groups together, facilitating the exchange of data. It provides a platform where data providers can list their datasets, and data consumers can browse, purchase, and download the data they need. Not only does this make the process of buying and selling data more efficient, but it also opens up opportunities for data providers to monetize their data and for data consumers to access a wider variety of data than they could on their own. In essence, a data marketplace democratizes data, making it more accessible and usable for a variety of purposes. It's an essential component of the data economy, where data is increasingly seen as a valuable resource.

Databricks Marketplace is an open marketplace for your data, analytics, and AI needs. It aims to address the challenges of data sharing, which is often hindered by technical, legal, and business demands, such as platform dependencies, data replication, security risks, and contractual agreements. Databricks Marketplace is powered by Delta Sharing and expands your opportunity to deliver innovation and advance your analytics and AI initiatives. In this chapter, you will learn how to navigate Databricks Marketplace and take advantage of its key benefits. You will explore topics such as understanding popular use cases and data providers across industry. You will also learn about the different types of data assets available on the Marketplace, including AI models and prebuilt notebooks and solutions, and how Delta Sharing and Marketplace work together. By the end of the chapter, you will have a comprehensive understanding of Databricks Marketplace and will also be able to apply best practices and tips for sharing data securely and efficiently with other organizations.

Benefits of Databricks Marketplace

Data providers and data consumers often face difficulties when it comes to sharing data. For example, existing data marketplaces that provide only datasets miss out on one of the key considerations for data consumers: the context around the data. Additionally, most current marketplaces work in walled garden environments. Data exchange can be done only on a marketplace's closed platform and sometimes only within their proprietary data formats. Databricks Marketplace is an open platform for data, analytics, and AI needs. It provides a secure and fluid data sharing capability powered by the open source Delta Sharing standard, offering a significant advantage over other marketplaces. It offers users the ability to discover, evaluate, and gain access to essential datasets, notebooks, and more, expanding opportunities for innovation while advancing analytics and AI initiatives.

One of the key features of Databricks Marketplace is its ability to provide consumers with access to datasets and to AI and analytics assets, such as ML models, notebooks, applications, and dashboards, without the need for proprietary platform dependencies, complicated ETL (extract, transform, load), or expensive replication. Delta Sharing makes this possible by allowing consumers to access data products without having to be on the Databricks platform. As a result, data providers can broaden their addressable market without forcing consumers into vendor lock-in. In addition to these benefits, Databricks Marketplace offers data products and AI assets from various providers, including free and paid options. Figure 3-1 illustrates the Marketplace UI within your Databricks workspace. Simply navigate to "Marketplace" to browse and search for both free and paid products across various categories and providers.



Open marketplace for data, analytics, and AI

Discover, evaluate, and gain access to Data sets, AI models, Solutions Accelerators, notebooks, and more...

Figure 3-1. Databricks Marketplace

Delta Sharing and the Marketplace

Fundamentally, Delta Sharing serves as the pillar of Databricks Marketplace. It fuels the Marketplace by allowing data providers to expand their potential market without imposing vendor restrictions on consumers. This positions Databricks Marketplace as a free platform for trading data products. By using the open source Delta Sharing standard, Databricks Marketplace enables consistent data sharing across organizations and platforms, reducing barriers to collaboration and innovation while making it easier for stakeholders to work together to achieve their goals. Delta Sharing and the Marketplace can be beneficial for you in several ways:

Simplified data sharing

Databricks Marketplace simplifies the sharing of data across organizations, streamlining the process for both data providers and consumers. It integrates with existing data workflows, enabling you to share their data with others or access data shared by others without the need for complex setup procedures.

Real-time data sharing
: Databricks Marketplace facilitates the sharing of live data in real time without the need for data replication. Data providers can share their data, ensuring the information is always up to date and readily available to consumers.

Open access and no vendor lock-in
: Databricks Marketplace promotes an open approach to data access, allowing you to connect to shared data using your preferred tools without being tied to the Databricks platform. This approach ensures flexibility and avoids vendor lock-in, allowing you to choose the tools that best suit your needs.

Secure and controlled data sharing
: Security is a paramount concern in Delta Sharing, which provides robust mechanisms for secure data sharing, ensuring that data is shared in a controlled and protected manner to instill confidence in the safety of shared data.

Centralized governance for shared data
: Databricks Marketplace offers centralized governance features for shared data. Data providers can control who can access their data and how it can be used, ensuring proper management and responsible utilization of shared resources.

Unity Catalog and the Marketplace

Unity Catalog and Databricks Marketplace are interconnected components within the Databricks ecosystem. Unity Catalog functions as a governance tool for data and AI assets, offering a unified location for managing and auditing data access. The link between the two emerges as Delta Sharing necessitates the activation of Unity Catalog in your Databricks workspace. When you utilize Marketplace data products through a Databricks workspace that has Unity Catalog enabled, you can benefit from the comprehensive integration between Delta Sharing and Unity Catalog. These benefits include governance by Unity Catalog, auditing capabilities, and user-friendly interfaces. Unity Catalog plays a pivotal role in controlling and overseeing the data shared through Databricks Marketplace. When combined with Databricks Marketplace, Unity Catalog offers several key benefits:

Unified visibility into data and AI with built-in auditing, discovery, and lineage

Unity Catalog provides a comprehensive solution for gaining insights into your data and AI assets. It enables you to discover and classify structured and unstructured data, ML models, notebooks, dashboards, and various files, irrespective of the cloud platform. The system automatically records user-level audit logs and captures lineage data, revealing how data assets are created and used across different programming languages. Additionally, Unity Catalog allows you to tag and document data assets, facilitating data discovery for users.

Single permission model for data and AI

You can simplify access management with a unified interface for defining access policies that apply to data and AI assets across all workspaces, centralizing access control for streamlined administration.

Standards-compliant security model

Unity Catalog's security model adheres to standard ANSI SQL, providing administrators with a familiar syntax for granting permissions within their existing data lake, thus enhancing security while ensuring ease of use.

AI-powered monitoring and observability

Unity Catalog allows you to utilize the capabilities of AI to automate monitoring, diagnose errors, and maintain the quality of data and ML models, improving the reliability and performance of your AI and data systems.

Open data sharing

You can effortlessly share data and AI assets across various clouds, regions, and platforms using open source Delta Sharing, integrated into Unity Catalog, allowing data sharing without constraints and enhancing collaboration across your organization.

The combination of Unity Catalog with Databricks Marketplace provides a powerful tool for managing, sharing, and discovering data products. This unified approach to governance accelerates data and AI initiatives while ensuring regulatory compliance in a simplified manner.

Industry Use Cases

You can exchange data products such as datasets, notebooks, dashboards, and machine learning models on the Marketplace, allowing data consumers to discover, evaluate, and access more data products from third-party vendors than ever before. Providers can commercialize new offerings and shorten sales cycles by providing value-added services on top of their data. The Marketplace can be used in various industries, such as healthcare, finance, retail, and manufacturing. Here are some examples of offerings by industry:

Healthcare and life sciences
: Access pretrained machine learning models for tasks like medical image analysis (e.g., X-rays and MRIs) and curated datasets for research and analysis in fields such as genomics and patient health records.

Financial services
: Utilize prebuilt financial analytics and risk models to conduct complex financial analysis and assess risks. Access historical financial datasets for backtesting trading strategies.

Retail and ecommerce
: Leverage machine learning models for demand forecasting and inventory optimization. Explore prebuilt customer analytics solutions to analyze behavior and preferences.

Manufacturing and industrial IoT
: Access predictive maintenance models for identifying potential equipment failures and optimizing maintenance schedules. Utilize real-time data streams to monitor and enhance manufacturing processes.

Media and entertainment
: Make use of natural language processing models for sentiment analysis and content recommendation. Access data sources that include social media trends and user engagement metrics.

Government and public sector
: Leverage prebuilt models for detecting fraud and anomalies in public programs. Access datasets related to demographics, public health, and environmental data for analysis and policy making.

Energy and utilities
: Access predictive models for energy demand forecasting and optimization. Utilize weather data and other relevant datasets for energy consumption analysis.

Transportation and logistics
: Utilize machine learning models for optimizing routes and managing fleets. Access datasets containing transportation and traffic data for analysis and decision making.

Marketplace Partners

Databricks Marketplace thrives on the contributions of its diverse partners. These partners, which include data providers, technology partners, and consulting partners, play a crucial role in enriching the Marketplace with a wide array of data products and services.

Data Providers

The Marketplace expands your ability to deliver innovation and advance your analytics and AI initiatives. It allows data consumers to discover, evaluate, and access more data products from third-party vendors than ever before. Providers can now commercialize new offerings and shorten sales cycles by providing value-added services on top of their data. There are hundreds of providers available across more than 15 industry categories. Here are some of the providers on the Marketplace that can benefit organizations in healthcare, finance, retail, manufacturing, and other industries:

LiveRamp
: Provides a privacy-conscious and configurable collaboration platform for organizations and external partners to create audiences, activate data, and access insights. This includes enhancing customer data with demographic and psychographic information from AnalyticsIQ, Catalina, Experian, Polk, and other sources.

AnalyticsIQ
: Offers demographic and psychographic data, including datasets for consumer demographics, lifestyle behaviors, and purchase intent.

Catalina

Specializes in personalized digital media solutions and provides datasets for purchase behavior insights.

Experian

Offers datasets such as credit scores and credit reports.

Polk

Provides automotive data solutions, offering a wide range of data products such as datasets for vehicle registration data.

ShareThis

Offers solutions to enhance digital marketing efforts, foster social engagement, and drive traffic to websites and digital properties for businesses.

Technology Partners

Technology partners integrate their solutions with Databricks to provide complementary capabilities for ETL, data ingestion, business intelligence, machine learning, and governance. These integrations enable you to leverage the Databricks Data Intelligence Platform's reliability and scalability to innovate faster while deriving valuable data insights.

Consulting Partners

Consulting partners are experts uniquely positioned to help you strategize, implement, and scale data, analytics, and AI initiatives with Databricks. They bring technology, industry, and use case expertise to help you make the most of the Databricks Data Intelligence Platform. Typical consulting partners include global, regional, and industry-leading consulting services and technology product companies.

Marketplace Datasets and Notebooks

Access to diverse, high-quality datasets is crucial for building robust and effective data science and machine learning models. Databricks Marketplace offers a wide range of datasets that cater to various domains and use cases. These datasets are made available through tables or notebooks that connect to live data via APIs. The notebooks can also include dashboards, data exploration scripts, and useful code snippets. There are currently almost five hundred datasets and notebooks, spanning more than 20 different categories and industries, that are available for free (instantly and by request) and through the private data exchange.

Table 3-1 introduces a selection of free datasets that are instantly available in Databricks Marketplace. These datasets, provided by several different organizations, span various categories and can be accessed either instantly or by request, making them accessible via Catalog Explorer within your workspace. Furthermore, each dataset has a brief description to help you understand the content and potential applications. There are also datasets available by request for which the data provider will review your request or reach out directly; you will need to provide your company name and intended use case.

Table 3-1. Free datasets available in Databricks Marketplace

Product name	Provider name	Category	Description
AutoIQ—automotive data	AnalyticsIQ	Advertising and marketing, Retail	Automotive preferences and behaviors ranging from the number of vehicles to those in the market for specific auto makes
FinanceIQ—consumer financial insights	AnalyticsIQ	Advertising and marketing, Financial, Retail	Individual- or household-level data variables that provide and predict a variety of consumer finance attributes
HealthIQ—health and wellness	AnalyticsIQ	Advertising and marketing, Demographics, Health	Individual- and household-level data that provides and predicts a variety of unique characteristics related to health
US Drinking Water System Ratings by Zipcode	aterio.io	Public sector, Health	Insights into drinking water system irregularities by the US EPA
US Housing Forecast by Zipcode and County - 2030	aterio.io	Demographics, Public sector, Health	Offers current housing data and vital insights into future housing needs, aligning with population growth trends
US Population Forecast by Zipcode and Counties - 2025,2030	aterio.io	Demographics, Geospatial	Offers insights into present demographics, historical population trends, and future projections
US Real Estate Investment Ratings by Zipcode	aterio.io	Demographics, Economics	The aggregate score provides a comprehensive overview of the suitability to buy a house in a particular zip code.

Product name	Provider name	Category	Description
IRS SOI	Claritype, Inc.	Financial, Public sector	IRS statistics of income (SOI) by state, county, and zip code
Property Characteristic Information on US Properties (Sample)	CoreLogic	Financial	Information on parcels, location, ownership, structure, legal status, and more for residential properties
Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)	Databricks	Health	This synthetically generated dataset contains records corresponding to 100,000 patients provided in OMOP5.3 format.

As an example, the COVID-19 - World Confirmed Cases, Deaths, and Testing dataset is made available by Rearc via a notebook. The dataset is a collection of COVID-19 data maintained by Our World in Data and will update daily throughout the duration of the COVID-19 pandemic. It includes vaccinations, test positivity rates, hospital and ICU admissions, confirmed cases and deaths, reproduction rate, policy responses, and other variables of interest. The goal of this notebook is to make the knowledge on big health problems related to COVID easily accessible and understandable. The notebook contains code for viewing and exploring the data as well as for performing pandas-based profiling reports through charts, correlations, and other visual reports.

Solution Accelerators

Solution Accelerators are prebuilt solutions designed to address common use cases and speed up the development of data and AI applications. They cover a wide range of domains, from cybersecurity to healthcare, and are developed by various providers, including Databricks and its partners.

Solution Accelerators offer a unique blend of data, models, and code, providing you with a solid starting point for your projects. They can help you reduce development time, avoid common pitfalls, and achieve better results.

Table 3-2 provides an overview of some of the Databricks Solution Accelerators available in the Marketplace.

Table 3-2. Databricks Solution Accelerators available in the Marketplace

Solution Accelerator	Description
Threat Detection with DNS	Uses DNS data to detect and investigate a malware campaign called Agent Tesla RAT
Real-Time Bidding Optimization	Uses BidRequest data to predict viewability and optimize media spend and ROI for marketing campaigns
Media Mix Modeling	Uses multiple data sources to measure and optimize the impact of marketing campaigns across different channels
Graph Analytics for Telco Customer Churn Prediction	Uses graph analytics to identify and predict which telco customers are likely to churn
Better LLMs with Better Data Using Cleanlab Studio	Uses Cleanlab Studio to improve the quality of training data and boost the performance of large language models
Indicator-of-Compromise (IOC) Matching	Uses Spark SQL and Delta Lake to perform fast and scalable IOC matching for security use cases
Accelerating Interoperability with Databricks Lakehouse	Uses FHIR data to enable interoperability and analytics for healthcare use cases
Digital Pathology Solution Accelerator	Uses deep learning to improve the efficiency and accuracy of diagnostic teams in pathology
Automated PHI Removal	Uses natural language processing to detect and protect sensitive patient data
Cyber Security Incident Investigation Using Graphistry	Uses Graphistry to visualize and analyze large-scale security data for incident response and threat hunting

Additionally, providers can include their own models and apps on the Marketplace for instant or request-based access, or through paid private exchanges. One example of a request-based "model" on Databricks Marketplace is LUCID, a trusted research environment that enhances the advanced analytics capabilities of TriNetX. It allows users to integrate data from TriNetX LIVE and Databricks and use an analytics toolkit developed by TriNetX. The toolkit includes two data formats (datasets and extracts), Python and R functions for data wrangling, and notebook templates for common analyses. LUCID enables users to perform more types of analyses on real-world data without downloading a dataset, increase transparency and flexibility of analysis methods, and power federated analytics and model deployment across TriNetX data networks.

AI Models

Databricks Marketplace enables easy access to ready-to-use AI models developed and provided by both Databricks and third parties, catering to a variety of use cases and domains. For instance, a provider could build a domain-specific natural language model to detect healthcare-specific clinical phrases. By offering AI models developed and provided by third parties, businesses can leverage these models for specific use cases and domains.

Individuals can contribute their AI models to Databricks Marketplace for utilization by others. Databricks Marketplace serves as an open platform for the exchange of data assets such as datasets, notebooks, dashboards, and AI models. Through AI model sharing, Databricks users can access state-of-the-art models that can be easily and securely implemented on their data.

To become a contributor of data on the Marketplace, one might need to participate in a partner program. This guarantees that the contributed data products, encompassing AI models, adhere to certain criteria and are appropriate for use by others.

Industry Use Cases

Databricks has published a curated list of open source models available within the Marketplace, including MPT-7B and Falcon-7B instruction-following and summarization models and Stable Diffusion for image generation, making it easy to get started with generative AI across a variety of use cases. Generative AI models can be used in a variety of industries to generate new content or data. Here are some examples of how generative AI models could be used in different industries:

Healthcare

Create synthetic medical data to support research efforts or develop tailored treatment plans for patients by analyzing their medical history and genetic data.

Finance

Generate synthetic financial data to enhance risk assessment and fraud detection capabilities or craft personalized investment strategies for clients in alignment with their financial objectives and risk preferences.

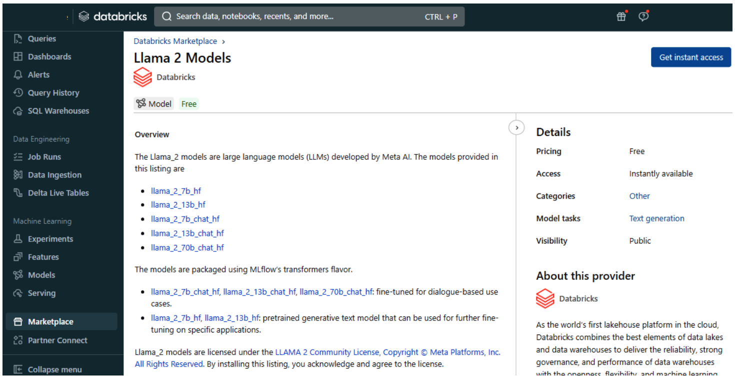
Retail
: Utilize synthetic data to inspire fresh product ideas and marketing materials or offer personalized shopping experiences by leveraging customer purchase history and preferences.

Manufacturing
: Enhance product design and streamline production processes using synthetic data or develop proactive maintenance schedules for equipment by analyzing historical data and usage patterns.

Getting Started with Llama 2 Models

Among the curated foundation models available on the Marketplace is the Llama 2 model family, developed by Meta AI. These models are large language models (LLMs) that have been trained on a vast amount of text data using deep learning techniques. They can generate humanlike text for a wide range of tasks, such as answering questions, summarizing documents, writing essays, and more.

The Llama 2 models available in this listing are the Llama-2-7b-chat-hf, Llama-2-13b-chat-hf, and Llama-2-70b-chat-hf models. These models are fine-tuned for dialogue-based use cases and can be used to generate responses to text-based instructions or for building chatbots and conversational AI systems. The models vary in size and capabilities, with parameters ranging from 7 billion to 70 billion. Figure 3-2 shows how easy it is to get instant access to Llama 2 models from Databricks Marketplace.



Llama 2 Models

Databricks

Overview

The Llama_2 models are large language models (LLMs) developed by Meta AI. The models provided in this listing are

- llama_2_7b_hf
- llama_2_13b_hf
- llama_2_7b_chat_hf
- llama_2_13b_chat_hf
- llama_2_70b_chat_hf

The models are packaged using MLflow's transformers flavor.

- llama_2_7b_chat_hf, llama_2_13b_chat_hf, llama_2_70b_chat_hf: fine-tuned for dialogue-based use cases.
- llama_2_7b_hf, llama_2_13b_hf: pretrained generative text model that can be used for further fine-tuning on specific applications.

Llama_2 models are licensed under the LLAMA 2 Community License, Copyright © Meta Platforms, Inc. All Rights Reserved. By installing this listing, you acknowledge and agree to the license.

Details

Pricing	Free
Access	Instantly available
Categories	Other
Model tasks	Text generation
Visibility	Public

About this provider

Databricks

As the world's first lakehouse platform in the cloud, Databricks combines the best elements of data lakes and data warehouses to deliver the reliability, strong governance, and performance of data warehouses with the openness, flexibility, and machine learning

Figure 3-2. Llama 2 models in Databricks Marketplace

To use these models, you need a Databricks workspace that is Unity Catalog enabled. Here is a summary of the steps to enable Unity Catalog in Databricks:

1. As an account admin, log in to the account console and go to “Workspaces.”
2. Turn on the “Enable Unity Catalog” option and choose the Metastore.
3. Confirm by clicking Enable and finish the workspace creation configuration by clicking Save.
4. After enabling Unity Catalog, you can create clusters or SQL warehouses for users to query and create objects. Don’t forget to grant privileges to users and create new catalogs and schemas as needed.

You also need to accept the terms and conditions of the Llama 2 Community License Agreement before installing the listing. Once it is installed, you can view detailed information about each model, including specifications, performance, limitations, and usage examples.

You can deploy these models directly to Databricks Model Serving for immediate use. This allows you to create REST endpoints for your models and serve them with low latency and high scalability. You can also load the models for fine-tuning or batch inference use cases using the MLflow API or the Transformers library.

Here’s how you can load the Llama-2-7b-chat-hf model using MLflow:

```
import mlflow
model_uri = "models:/Llama-2-7b-chat-hf/1"
model = mlflow.pyfunc.load_model(model_uri)
```

And here’s how you can load the same model using Transformers:

```
from transformers import AutoTokenizer, AutoModelForCausalLM
model_name = "meta-ai/llama-2-7b-chat-hf"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)
```

Once loaded, you can use the model to generate texts based on your inputs—for example:

```
input_text = "Hello, how are you?"
input_ids = tokenizer.encode(input_text + tokenizer.eos_token,
return_tensors="pt")
output_ids = model.generate(input_ids)
output_text = tokenizer.decode(output_ids[0], \
    skip_special_tokens=True)
print(output_text)
```

This will generate a response to the greeting "Hello, how are you?"

Another technical capability of Delta Sharing is its open sharing connectors, which are libraries that allow data consumers to access shared data using various tools and frameworks, such as pandas, Apache Spark, Rust, and more. To access shared data, data consumers need to obtain a shared credential from the data provider, which is a JSON file that contains the necessary information. Data consumers also need to install the appropriate connector for their tool or framework and use it to read the shared data. However, there are some limitations to this data sharing mechanism. Data consumers can only read the shared data and cannot write or modify it. Data consumers also need to handle data encryption and decryption themselves if the shared data is encrypted at rest. Data consumers may also encounter performance issues or errors depending on the size and complexity of the shared data.

Getting Started

To start consuming data from the Marketplace, you must first have a Premium Databricks account and workspace, a Unity Catalog metastore, and the `USE MARKETPLACE ASSETS` privilege. If your admin has disabled this privilege, you can request that they grant it to you or grant you either the `CREATE CATALOG` or the `USE PROVIDER` permissions on the Unity Catalog metastore. If you do not have any of these privileges, you can still view Marketplace listings but cannot access data products.

Once you have an account and the relevant permissions, you can browse or search for the data product you want on Databricks Marketplace. You can filter listings by provider name, category, cost (free or paid), or keyword search. Since Databricks Marketplace uses Delta Sharing to provide security and control over shared data, consumers can access public data, free sample data, and commercialized data offerings. In addition to datasets, consumers can leverage additional analytical assets such as Databricks notebooks to help kick-start the data exploration process. Databricks offers a pay-as-you-go approach with no up-front costs. You pay only for the products you use at per-second granularity. Some data products are free, while others may have a cost associated with them. Once you have access to the data products, you can access them via Data Explorer, Databricks CLI, or SQL statements. You would also be able to grant other users access to the catalog that contains the shared data and transfer ownership of the catalog or the objects inside it.

For data providers, Databricks Marketplace gives a secure platform for sharing data products that data scientists and analysts can use to help their organizations succeed. Providers can share public data, free sample data, and commercialized data offerings. They can also share Databricks notebooks and other content to demonstrate use cases and demonstrate how to take full advantage of their data products.

To create a data product in the Marketplace, you'll need a Premium Databricks workspace with access to tools and services to prepare your data, such as Delta Lake, MLflow, or SQL Analytics. You can also use external tools or libraries that are compatible with Databricks. You'll then need to package your data product using one of the supported formats, such as Delta Sharing or MLflow Model Registry. To list your data products on the Marketplace, you'll need to apply to be a provider through the Databricks Data Partner Program and review the Marketplace provider policies. You can then publish your data product in the Marketplace using the Publish Data Product UI in your workspace. You will need to provide some information about your product, such as name, description, category, price, terms, and conditions. You will also need to agree to the Databricks Data Provider Agreement.

Deploying and serving a model from the Marketplace follows a similar process. You'll need to use Databricks Model Serving. This is a feature that allows you to create REST endpoints for your models and serve them with low latency and high scalability. You can access Databricks Model Serving from the left sidebar menu in your workspace. You can then select a model from Catalog Explorer or MLflow Model Registry and click on Deploy Model. You will then need to configure some settings for your model deployment, such as name, version, inference cluster size, and so on. Once your model is deployed, you will get a URL for your model endpoint that you can use to make predictions.

Becoming a Data Provider Partner with Databricks offers several benefits. It allows you to reach a larger audience of data consumers through a single secure platform, enhancing the customer experience by reducing setup and activation time. Databricks also provides marketing support to increase your exposure. As a partner, you can leverage Databricks's market-leading Data Intelligence Platform for data, analytics, and AI. You also gain access to Databricks's product, engineering, and support staff. Furthermore, you can collaborate with Databricks's industry teams to build industry-specific solutions designed for various customer use cases.

Summary

In this chapter, you learned about the advantages of Databricks Marketplace for data collaboration. Powered by Delta Sharing, the Marketplace brings the capability of open cross-platform and live data sharing without data replication. Additionally, the centralized governance model provides added security and control. Databricks Marketplace ensures secure and compliant data sharing between data providers and consumers. Data providers can control who can access their data products, set terms and conditions, and monitor usage and billing. Data consumers can trust that the data products are verified and validated by Databricks and the providers. The Marketplace also leverages Databricks's built-in security and governance features, such as encryption, authentication, authorization, auditing, and data masking. In addition, the Marketplace supports privacy-preserving technologies, such as differential privacy and federated learning, to protect sensitive data.

From empowering business intelligence with actionable insights to fueling data science and machine learning endeavors, you discovered how data integration, governance, storytelling, and data art can be enriched through this platform. With a range of custom datasets, notebooks, and AI models becoming available within the Marketplace, digital transformation and innovation initiatives can be pursued within the Databricks platform. In Chapter 4, you will learn more about safeguarding data privacy with Databricks Clean Rooms and other essential practices to ensure that sensitive data remains secure and inaccessible to unauthorized entities.

CHAPTER 4

Safeguarding Data with Clean Rooms

With the rise of data privacy regulations such as the GDPR and the CCPA and the increasing demand for external data sources, such as third-party data providers and data marketplaces, organizations need a secure, controlled, and private way to collaborate on data with their customers and partners. However, traditional data sharing solutions often require data replication and trust-based agreements, which expose organizations to potential risks of data misuse and privacy breaches.

The demand for data clean rooms has been growing in various industries and use cases due to the changing security, compliance, and privacy landscape, the fragmentation of the data ecosystem, and the new ways to monetize data. According to Gartner, 80% of advertisers that spend more than $1 billion annually on media will use data clean rooms by 2023.[1] However, existing solutions have limitations on data movement and replication, are restricted to SQL, and are hard to scale.

The Databricks Data Intelligence Platform provides a comprehensive set of tools to build, serve, and deploy a scalable, flexible, and interoperable data clean room based on your data privacy and

[1] Interactive Advertising Bureau (IAB), *State of Data 2023: Data Clean Rooms and the Democratization of Data in the Privacy-Centric Ecosystem*, January 24, 2023, *https://oreil.ly/OBVgm*.

governance requirements. Some of the features include secure data sharing with no replication, full support to run arbitrary workloads and languages, easy scalability with guided onboarding experience, isolated compute, and being privacy-safe with fine-grained access controls.

Databricks Clean Rooms enable organizations to share and join their existing data in a secure, governed, and privacy-safe environment. Participants in the Databricks Clean Rooms can perform analysis on the joined data using common languages such as Python and SQL without the risk of exposing their data to other participants. Participants have full control of their data and can decide which participants can perform what analysis on their data without exposing sensitive data, such as personally identifiable information (PII).

This chapter provides an in-depth look at how Databricks Clean Rooms work and how they can help organizations guard their data privacy. The chapter also explores key partnership integrations that enhance the capabilities of Databricks Clean Rooms and provide additional benefits for data privacy and security. By using Databricks Clean Rooms with these partner solutions, organizations can unlock new insights and opportunities from their data while preserving data privacy.

Challenges and Solutions for Safeguarding Data

Implementing and using Databricks Clean Rooms can present several challenges, but Databricks provides solutions to address these issues effectively:

Data privacy and security
: One of the main challenges is ensuring data privacy and security. When multiple participants join their first-party data and perform analysis, there is a risk of exposing sensitive data to other participants. Databricks Clean Rooms provide a secure, governed, and privacy-safe environment.

Data standardization
: Several data clean rooms have not yet adopted universal standards for their implementation. This means that platforms and advertisers may be trying to pool data that exists in multiple

formats, making the prep work for aggregating those different formats time-consuming. Databricks Clean Rooms allow computations to be run in any language, including SQL, R, Scala, Java, and Python, which enables simple use cases such as joins as well as complex computations such as machine learning, supporting data in multiple formats.

Scalability

As the number of participants increases, it becomes more difficult to manage the clean room environment. Databricks Clean Rooms are designed to easily scale to multiple participants. They also reduce time to insights with predefined templates for common clean room use cases.

Interoperability

Interoperability is the ability of different systems, devices, or applications to exchange and use data without requiring special adaptations or conversions. Interoperability is a significant challenge when working with data across different clouds, regions, and platforms. With Delta Sharing, clean room collaborators can work together across clouds, across regions, and even across data platforms without requiring data movement. By addressing these challenges, Databricks Clean Rooms enable businesses to collaborate securely on any cloud in a privacy-safe way.

Databricks Clean Rooms Explained

Databricks Clean Rooms are innovative, closed-loop environments designed to facilitate seamless collaboration among disparate parties, all while safeguarding sensitive information and preserving proprietary data. By providing a secure and privacy-safe haven for data sharing, Clean Rooms emerge as a pivotal asset for organizations navigating intricate data-driven endeavors.

As organizational data flows from various sources, including media platforms, walled gardens, and collaborative partners, Clean Rooms offer a sanctuary in which data privacy remains unwavering. They cater to the unique needs of data-driven organizations seeking to unlock insights from diverse sources within a fragmented and regulated data ecosystem. Clean Rooms usher in a host of compelling advantages, each contributing to a robust and efficient data collaboration framework. They grant unparalleled access to data and intellectual property, fostering an environment in which partners can collaborate and innovate effortlessly, which results in expanded avenues for partners and enables them to explore new use cases and automated workflows. The gains ripple further, bringing heightened efficiency, productivity, and scalability—indispensable attributes for organizations operating at scale. Notably, Clean Rooms are pivotal in safeguarding existing investments, acting as custodians of valuable data assets in an ever-evolving digital landscape.

The applications of Clean Rooms span a multitude of industries, each finding its unique utility. From optimizing campaign performance and enhancing personalization in consumer-centric fields to curbing fraud and mitigating risk in the financial sector, Clean Rooms empower data-driven decision making.

Constructing and nurturing Clean Rooms is a methodical process, orchestrated through the establishment of secure data connections across cloud platforms, the utilization of containerized code to fuel diverse use cases, and the defining of roles and permissions for collaborators. The foundation of privacy is fortified through the application of privacy-enhancing technologies (PETs), including encryption, obfuscation, data minimization, differential privacy, and noise injection. Stringent policies further ensure that data usage adheres to approved queries only, fostering a controlled and privacy-conscious environment.

Clean Rooms promote interoperability and automation by encouraging collaboration across multiple clouds and platforms without necessitating data movement. This environment provides a unified user experience, supports templatized analytics and natural language queries, simplifies user interactions, and facilitates ease of use. Realizing the potential of Clean Rooms necessitates a well-defined strategy and adherence to best practices. This journey encompasses pivotal steps such as data auditing, stakeholder education, sandbox testing, partner onboarding, ongoing operations management, preparation for data science tasks, centralization of outputs, and dissemination of insights.

Now available in private preview, Databricks Clean Rooms can effectively compartmentalize data, fostering collaborative workflows that propel innovation while steadfastly upholding privacy and regulatory compliance. In an era in which the demand for external data soars, Clean Rooms emerge as a secure data exchange, enabling organizations to embrace external data sources with confidence and drive data-driven evolution.

Crossing industries, Clean Rooms redefine possibilities. Consumer packaged goods (CPG) companies can harness the synergy of first-party advertisement and point-of-sale (POS) transactional data for sales uplift. The media industry could see a new era of targeted advertising, enhanced segmentation, and transparency in ad effectiveness, all while preserving data privacy. In financial services, the value chain aligns for proactive fraud detection and anti-money-laundering strategies. As organizations strive to balance innovation and compliance, Clean Rooms will secure data collaboration, shaping a future in which insights flow freely and privacy remains persistent.

Key Partnership Integrations

Databricks Clean Rooms are strengthened by key partnerships with industry-leading companies like Habu, Datavant, LiveRamp, and TransUnion. These partnerships are crucial pillars that elevate the data privacy and security capabilities of Databricks Clean Rooms. Let's look at a few of these partnerships that are enabling enhanced data-driven insights:

Habu

Habu offers a software platform that empowers brands to construct clean rooms alongside their partners, extending the ability to measure marketing campaign impact across diverse channels and platforms. The uniform integration with Databricks amplifies user experience and cross-platform interoperability, opening avenues for insightful analyses that uphold data privacy standards. Habu offers a simple and intuitive interface, supports data sharing across different clouds and platforms, and enables privacy-preserving analytics with Databricks.

Datavant

Datavant offers innovative tokenization technology, a boon for healthcare organizations seeking to unleash the potential of data-driven healthcare analysis. The partnership provides a transformative capability through the integration of Datavant tokens within a Databricks clean room, allowing data to be linked and analyzed comprehensively without compromising patient privacy and regulatory compliance.

LiveRamp

The LiveRamp integration accentuates the importance of secure data connectivity. A data connectivity platform, LiveRamp enables media entities and advertisers to harness their data assets across the digital landscape without sacrificing data privacy, bolstering its capacity to enable effective advertising targeting while upholding the sanctity of data privacy. It allows users to share and analyze data across different clouds and platforms, while maintaining data privacy and compliance. It also enables users to perform advanced analytics, such as machine learning, on data from multiple sources.

TransUnion

TransUnion, a global information and insights powerhouse, adds a layer of risk assessment and decision-making capabilities to Databricks Clean Rooms. The integration with TransUnion enriches the arsenal of tools available to businesses, equipping them with accurate and insightful information for informed decision making, all the while safeguarding data privacy.

Collectively, these pivotal partnerships further strengthen the data privacy and security capabilities of Databricks Clean Rooms. By fusing cutting-edge technologies with a commitment to data protection, businesses can make impactful decisions, fortified by the assurance of stringent data privacy standards. In the pursuit of excellence in today's data-centric landscape, Databricks Clean Rooms serve as trusted platforms. These platforms, strengthened by strategic partnerships, facilitate innovation while maintaining stringent data protection standards.

Industry Use Cases

Databricks Clean Rooms offer a wide range of applications across various industries. Their strong governance and data privacy enable collaborative data exploration in a secure environment, allowing multiple stakeholders to leverage their first-party data for analysis, while protecting proprietary information and upholding data privacy. The utility of Databricks Clean Rooms is demonstrated through diverse industry use cases, each showcasing the power of data sharing and collaboration within a framework that adheres to strict data privacy protocols:

Retail pioneering

The retail domain thrives on unified collaboration between retailers and suppliers, and Databricks Clean Rooms are reshaping the retail landscape by serving as secure conduits for confidential information exchange, underpinning demand forecasting, inventory planning, and supply chain optimization, elevating product availability, streamlining operations, and yield cost efficiencies.

Healthcare's data nexus

In the healthcare sector, Databricks Clean Rooms are custodians of sensitive healthcare data. Collaborators seamlessly meld and query diverse data sources, culminating in a nuanced understanding for real-world evidence (RWE) applications. From regulatory decision making to clinical trial design and observational research, data privacy remains sacred, fostering an environment of ethical and secure innovation.

Media's secure haven

Databricks Clean Rooms are catalyzing a transformative shift in the media industry by enabling secure sharing of audience data among media companies, advertisers, and partners, allowing for comprehensive analysis without infringing on user privacy and opening a new realm of collaborative insights, all within the confines of stringent data privacy regulations.

Financial integrity

In financial services, Databricks Clean Rooms align with the stringent Know Your Customer (KYC) standards. This collaboration augments the fight against financial malfeasance, facilitating comprehensive transaction investigations through collaborative analytics. In these secure environments, a holistic view of transactions materializes, allowing financial entities to address challenges with rigor and precision.

Driving automotive innovation

Databricks Clean Rooms are fostering synergy in the automotive industry by enabling secure data collaboration between manufacturers and suppliers, serving as crucial hubs for confidential data exchange and driving collaborative efforts in product development and supply chain optimization, all while ensuring the integrity of data.

Telecommunications unveiled

In the telecommunications sector, new paths of collaboration are being unlocked as carriers and service providers come together within secure environments to fuel cooperative efforts in optimizing networks and enhancing customer experiences, all while preserving the integrity of data privacy.

Public sector cohesion

Government agencies benefit from Databricks Clean Rooms by fostering a secure platform for data exchange with private sector counterparts, powering policy development and elevating service delivery within a secure and guarded environment.

Empowering energy

Energy companies and regulators connect within Databricks Clean Rooms to utilize secure data sharing for collaborative pursuits in energy production and distribution, steering the industry toward a sustainable future while upholding the sanctity of data privacy.

Educational advancement

Databricks Clean Rooms redefine educational collaboration, as academic institutions securely exchange insights with peers and government bodies, nurturing education policy development and service delivery while upholding the tenets of data privacy.

Organizations across various sectors are harnessing the power of data exploration, brought together by secure platforms. As they navigate this terrain, their commitment to the highest data privacy standards remains steadfast. This journey is steering innovation toward a future in which insights are seamlessly integrated and data protection is of the utmost importance.

Implementing Clean Rooms

In this section, you will learn about the details related to the implementation of Databricks Clean Rooms and how it provides a comprehensive set of tools to build, serve, and deploy a data clean room based on your data privacy and governance requirements. Before setting up a data clean room, clearly define your objectives and use cases. Ensure that you classify and segment your data based on sensitivity levels, access requirements, and compliance considerations. Also, plan for implementing robust security measures to protect the data clean room and the sensitive data it houses. The following section describes the key steps to implement Databricks Clean Rooms:

1. *Set up the Databricks Data Intelligence Platform.*

   The first step in implementing a Databricks clean room is to set up the Databricks Data Intelligence Platform, which provides a comprehensive set of tools to build, serve, and deploy a scalable and flexible data clean room based on your data privacy and governance requirements.

2. *Create a clean room and invite participants.*

   Next, create a clean room within the Databricks Data Intelligence Platform and specify the clean room participants. The isolated environment in which all jobs are executed is auto-created; no collaborator will be able to access the workspace for privacy/security reasons.

3. *Share data.*

   Participants can share their data with the clean room by uploading it to the Databricks Data Intelligence Platform. Data is not stored in the clean room but is instead shared into the clean room. No users have direct access to the datasets. They can access them only through approved notebooks/code. Data sharing among collaborators is secure and private. Only table metadata is visible to collaborators, while raw data remains inaccessible and hidden.

4. *Run computations.*

   Once data has been shared and access controls have been set up, participants can run computations on the data within the clean room. Computations can be performed using any language—SQL or Python—enabling simple use cases such as joins and crosswalks as well as complex computations such as machine learning.

Best Practices

In addition to setting up access controls within the clean room to ensure that only authorized users can access and process data, you should observe these best practices for using Databricks Clean Rooms to ensure data privacy and security:

Monitor data usage.

Monitor the usage of data within the clean room to ensure that the data is being used in compliance with data privacy and security policies. This can be done using tools such as audit logs and data usage reports.

Encrypt data.

Encrypt data at rest and in transit to ensure that it is protected from unauthorized access. This can be done using encryption tools provided by the Databricks Data Intelligence Platform or third-party encryption tools.

Implement data retention policies.

Implement data retention policies to ensure that data is not retained for longer than necessary. This can help to minimize the risk of data breaches and ensure compliance with data privacy regulations.

Regularly review and update security measures.

Regularly review and update security measures to ensure that they are effective in protecting data privacy and security. This can include updating access controls, monitoring tools, encryption tools, and data retention policies.

Define clear data sharing policies.

Define clear data sharing policies that outline the terms and conditions under which data can be shared within the clean room. This can help to ensure that all participants understand their rights and responsibilities when it comes to data sharing.

Provide training and support.

Provide training and support to clean room participants to help them understand how to use the clean room effectively. This can include training on how to share data, how to set up access controls, and how to run computations on the data within the clean room.

Leverage partner integrations.

Take advantage of partner integrations such as Habu, Datavant, LiveRamp, and TransUnion to enhance the capabilities of your Databricks Clean Room. These partners provide tools and technologies that can help you to improve data privacy and security within the clean room.

Regularly review and update data sharing agreements.

Regularly review and update data sharing agreements with clean room participants to ensure that they remain relevant and up to date. This can help to ensure that data sharing within the clean room remains compliant with data privacy regulations.

Future Trends

Several trends are likely to shape the evolution of clean room technology and the broader landscape of data privacy and collaboration:

Ubiquitous adoption of clean rooms

The adoption of clean room technology is expected to grow significantly. With cloud hyperscalers introducing clean rooms (such as AWS Clean Rooms from Amazon) alongside database companies like Snowflake and Databricks, the technical capability to perform a secure, double-blinded join is now accessible to engineers and product builders familiar with these stacks. As a result, you can expect to see clean-room-powered data flows emerging in most data-driven applications across myriad industries, from advertising to healthcare.

Enhanced walled garden solutions

A wave of new product developments and enhancements can be expected in the walled garden clean rooms, such as Google's Ads Data Hub, Amazon's Marketing Cloud, and Meta's Advanced Analytics. By enhancing advanced federated learning techniques, multiple data owners could share their data and AI assets and models without exchanging their data. Instead, they would share only the model updates or parameters, which are aggregated and applied to the global model. This way, the data remains local and private, while the model benefits from the collective data.

Increased interoperability

As more businesses adopt clean rooms, there will be a growing need for interoperability among different clean rooms across different clouds, regions, and platforms, driving further innovation in clean room technology to ensure seamless collaboration across different environments.

Greater focus on privacy

With increasing regulations around data privacy and the upcoming demise of third-party cookies, the scale and breadth of data sharing is becoming increasingly limited, leading to an increased focus on privacy-preserving technologies within clean rooms. Clean rooms could integrate advanced privacy techniques such as *differential privacy*, which adds controlled noise to data queries or outputs to protect the individual privacy of data records. It ensures that the statistical results of data analysis do not reveal any information about specific individuals in the data.

Enriched encryption

Clean Rooms can benefit from advanced encryption techniques such as *homomorphic encryption*, a technique that allows data owners to perform computations on encrypted data without decrypting it. This enables data analysis and machine learning on encrypted data without compromising data security or privacy.

Rise of orchestration

As data clean rooms become more complex and involve more participants, there will be a growing need for orchestration tools to manage these environments effectively.

The future of clean room technology looks promising, with several exciting trends on the horizon. As businesses continue to navigate the challenges of data privacy and collaboration, clean rooms will play an increasingly important role in enabling secure and effective data analysis.

Summary

This chapter provided an understanding of Databricks Clean Rooms, an innovative technology designed to protect data while enabling effective collaboration and analysis. The concept of clean rooms was explained, illustrating how they offer a secure, governed, and privacy-safe environment for data analysis. You learned about key partnerships that enhance the capabilities of Databricks Clean Rooms, along with various industry use cases that demonstrate the versatility and applicability of clean rooms across sectors. You also learned about steps for getting started with and implementing Databricks Clean Rooms, along with best practices for using clean rooms effectively. Clean rooms have the potential to evolve by integrating advanced techniques, such as differential privacy, federated learning, or homomorphic encryption, for robust data privacy and security. They could also provide more granular and dynamic control over data access and usage, enabling participants to modify their data sharing policies in response to changing business needs or regulatory requirements. Furthermore, they could support more complex and collaborative scenarios.

Databricks Clean Rooms represent a significant advancement in data privacy and collaboration. They enable businesses to leverage their data effectively while ensuring compliance with privacy regulations. As you navigate the evolving landscape of data privacy, clean rooms will play an increasingly important role in enabling secure and effective data analysis. This makes it an exciting area to watch for anyone interested in data privacy and collaboration. In Chapter 5, you will shift your focus to the strategic aspect of data collaboration, and you'll be guided through the process of developing an effective strategy for data collaboration.

CHAPTER 5

Crafting a Data Collaboration Strategy

Data collaboration is a multifaceted process that involves the sharing, integration, and analysis of data and AI assets by multiple entities—individuals, teams, or organizations—to achieve shared objectives. This dynamic process extends beyond the simple exchange of datasets and leverages shared data and AI assets for collaboration, creating knowledge and value.

A robust technical platform that can handle large volumes of data, support various data formats, provide advanced analytics capabilities, and facilitate seamless integration with other systems and tools is critical for enabling a unified and collaborative data environment.

Clear governance structures should define roles, responsibilities, and access rights, ensuring all participants understand their part in the data collaboration process. This clarity promotes accountability and helps prevent misuse of data.

Respect for privacy is paramount. A collaborative data strategy should provide mechanisms to protect sensitive information, including features like anonymization, secure data sharing, and lineage tracking.

Trust among participants can be fostered through transparency, open communication, and a shared understanding of the benefits of data collaboration. Regardless of the level of trust among collaborators, it is paramount that the appropriate tools are in place to manage and enable collaboration, including open and secure data sharing, private data exchanges, and privacy-safe clean room environments.

Crafting an effective data collaboration strategy offers numerous advantages for businesses, which can gain insights, accelerate innovation, and better position themselves to compete in the digital economy. It accelerates innovation by providing diverse perspectives and insights, leading to more inventive and effective solutions. It enhances data quality by allowing multiple parties to enrich the data and unlock value by creating new data products and applications that were not feasible in isolation.

In an increasingly interconnected world, the focus is shifting toward eradicating organizational and geographical barriers. This approach fosters a seamless collaboration environment, irrespective of the location or structure of the teams involved. Tools such as Delta Sharing, private exchanges, and clean rooms are instrumental in this process, enabling efficient collaboration across different trust levels. This is not just about a specific platform or tool but about a broader perspective on enhancing global collaboration.

Organizations can derive new value and accelerate innovation from their existing data assets through data collaboration. By sharing and integrating data, organizations can discover new insights, develop new services, and generate new revenue streams.

In this chapter, you will learn about different scenarios, technologies, architectures, and best practices for data collaboration that are possible with Databricks Delta Sharing. You will explore how Delta Sharing's open source approach can help you overcome the challenges and unlock the benefits of data collaboration across different clouds, platforms, and regions. You will also learn how to design and implement a data collaboration framework that aligns with your goals and scope, partners and roles, agreements and rules, platform and tools, and security and governance. You will also discover how to measure the success of your data collaboration strategy and how to manage change and improvement within your organization.

By the end of this chapter, you will have a comprehensive understanding of data collaboration with Databricks Delta Sharing and how it can transform how you create, access, use, and share data. You will also be able to apply the concepts and techniques learned in this chapter to your data collaboration projects and use cases.

Data Sharing Scenarios

Databricks Delta Sharing serves as a cornerstone for data collaboration, enabling seamless interaction within and across enterprises, cloud regions, and providers. This section delves into several scenarios for providers and recipients and into the technologies that underpin effective data sharing, including Databricks Delta Sharing, Unity Catalog, Databricks Marketplace, and third-party data platforms and tools. The aim is to provide a comprehensive understanding of data sharing scenarios and of how Delta Sharing can effectively enhance data collaboration. Whether you want to share data within your organization or with external partners and customers, Delta Sharing has you covered. For example, you can use Delta Sharing to share data between different departments in your organization that may be spread across different regions, such as marketing, sales, and research and development. In this way, you can improve coordination and decision making across your organization. You can also use Delta Sharing to share data with external parties, such as suppliers, distributors, or regulators, who may store their data across different regions, clouds, or even platforms. By enabling data sharing, you can streamline operations and compliance with your partners and customers seamlessly and effectively.

Delta Sharing leverages several technologies to facilitate efficient and secure data sharing. For instance, Delta Sharing servers handle the data sharing, ensuring that only authorized recipients can access the data. Unity Catalog provides a unified view of all available shares, making it easy for recipients to discover and access the data they need. Delta Sharing also integrates with third-party data platforms and tools, allowing recipients to access shared data using their preferred tools. For example, a data analyst using Tableau or Power BI for data visualization can access a share via Delta Sharing and directly import the data into Tableau or Power BI for analysis.

Internal Line of Business Sharing

Delta Sharing enables seamless data sharing across different departments within an organization. For instance, the marketing team can share customer insights with the sales team to help the latter tailor its strategies. Similarly, the R&D team can share product data with the marketing team for market analysis. This internal line of business sharing fosters better collaboration and decision making across the organization.

Consider a multinational corporation with departments spread across different geographical locations, as shown in Figure 5-1. Each department might have its own data lake in which to store its data. For instance, the sales department in the US might have sales data stored in an AWS S3 bucket, while the marketing department in Europe might have marketing data stored in an Azure Blob Storage.

With Delta Sharing, you can create a share of the sales data on the Delta Sharing server running in the AWS region (see Figure 5-1). The marketing department can then access this share using its preferred data analysis tool, such as Databricks Notebooks or Tableau, running in the Azure region. This is achieved without the need for data movement, thereby saving costs and reducing latency.

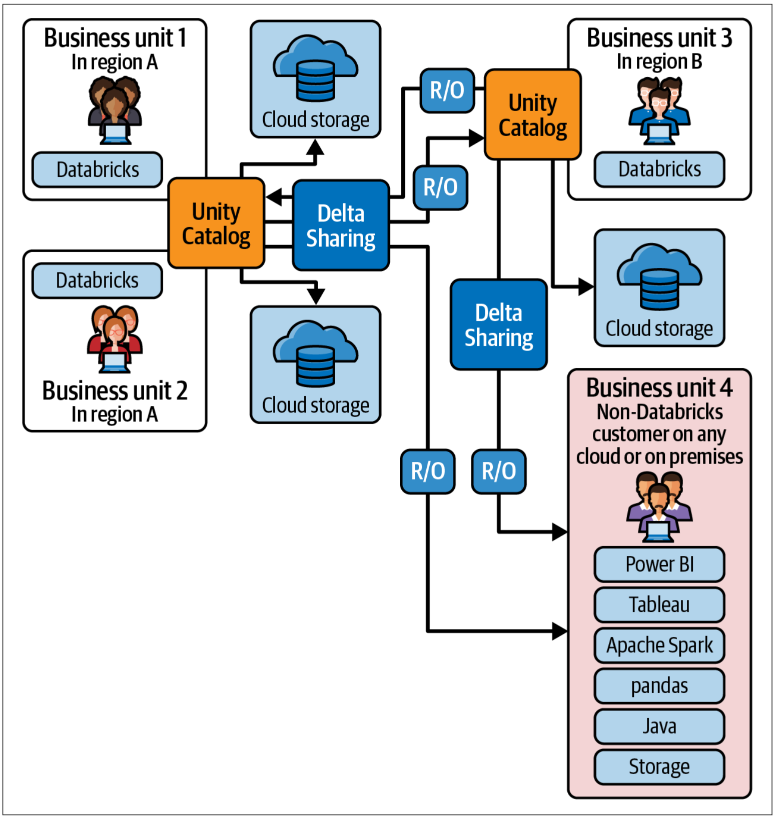


Figure 5-1. An example of Delta Sharing running in the AWS region

Business-to-Business Sharing

Delta Sharing also facilitates business-to-business (B2B) data sharing. Companies can share data with their business partners, such as suppliers, distributors, or service providers, to streamline operations. For example, a manufacturing company can share production data with its suppliers to ensure timely delivery of materials. Similarly, a retail company can share sales data with its distributors to manage inventory effectively.

A common scenario that arises is when a business is sharing data with its customers and needs to collect and merge data from various sources to create a unified dataset, as depicted in Figure 5-2. Such datasets can be critical for specific business needs, including decision making, market analysis, and research, and for supporting overall business operations.

Note that you can set up using change data feed (CDF) on a shared table across all three clouds to enhance performance and reduce egress costs. Then, within each cloud region, data can be shared with the customers. However, a newer approach uses Cloudflare R2 as an external storage location to be more efficient and cost-effective. Cloudflare R2 is in public preview at the time of this writing.

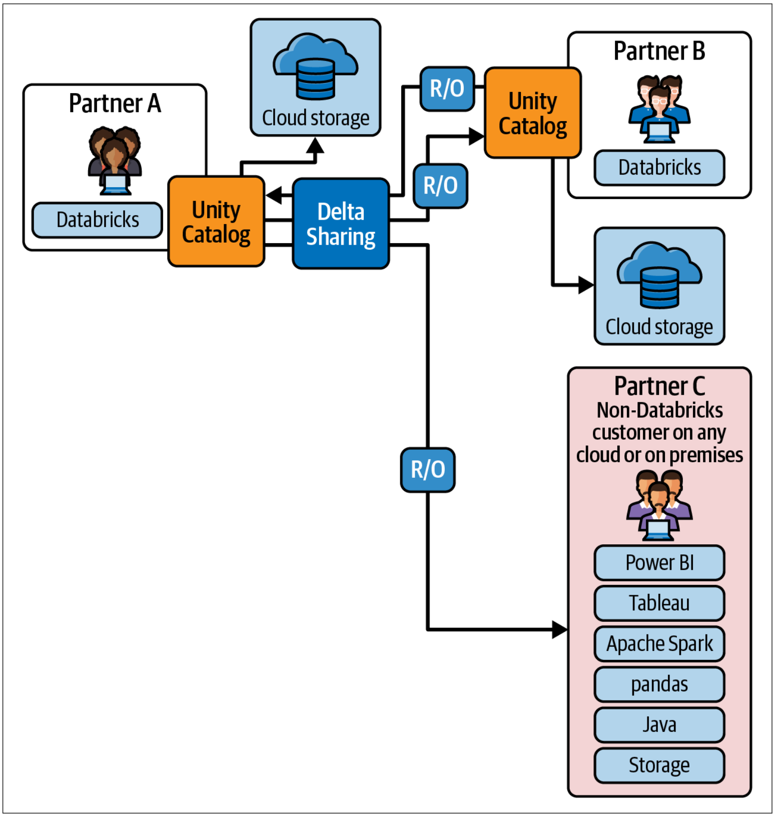


Figure 5-2. An example of business-to-business sharing in which a business needs to collect and merge data from various sources to create a unified dataset

Data Monetization

Data monetization is another key scenario in which Delta Sharing can be leveraged. Companies can monetize their data by sharing it with external parties, as illustrated in Figure 5-3. For instance, a telecom company can share user behavior data (while ensuring privacy and compliance) with marketing agencies for targeted advertising. Similarly, a financial institution can share market trends with investment firms for a fee. Delta Sharing ensures the data sharing process is secure, efficient, and compliant with regulations.

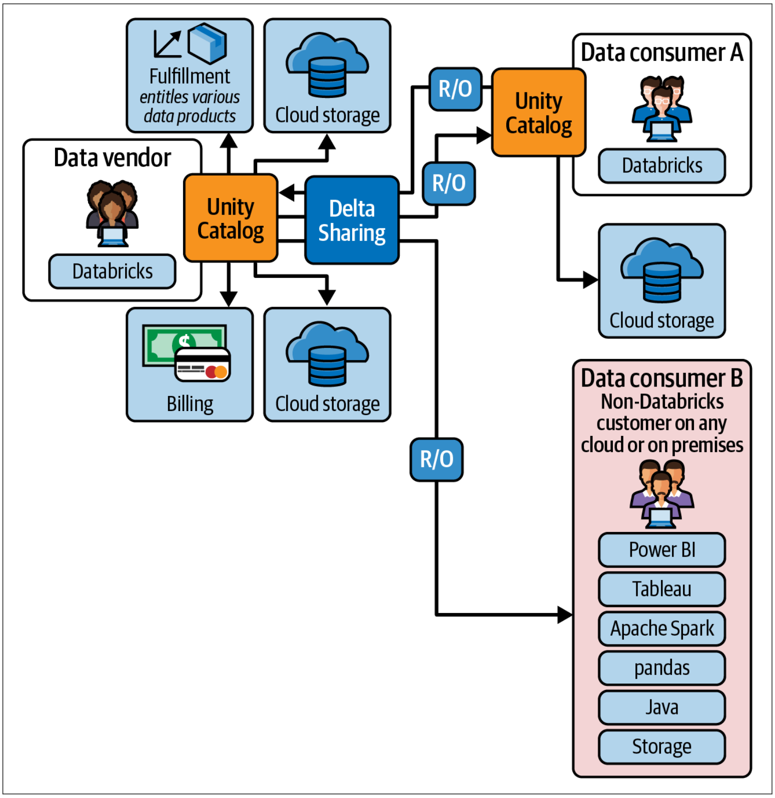


Figure 5-3. Delta Sharing allows companies to share data with partners securely and efficiently, for a fee

Designing a Data Collaboration Framework

Data collaboration has the potential to unlock new insights, drive innovation, and create value from data. However, it also presents a unique set of challenges, including data governance, privacy, security, trust, and interoperability. A well-structured strategy is essential to navigating these complexities and harnessing the power of data collaboration.

Designing a data collaboration strategy is a process that requires careful consideration of various elements. It's not just about the technology or platform being used but about how these tools and features support the overall framework and strategy.

The process begins with secure data sharing. An open protocol for secure data sharing, such as Delta Sharing, can be a game changer. It allows for the real-time exchange of large datasets in a safe and controlled manner, facilitating collaboration across different organizations without compromising security or privacy.

Next, the strategy must ensure a secure and compliant environment for data analysis. Features such as clean rooms can be instrumental in this regard, especially in healthcare and other sectors in which data privacy is paramount.

A unified data catalog, akin to the Unity Catalog, can enhance data governance and make data management more efficient. It provides a single, collaborative, and governed workspace for all your data, streamlining the process of data collaboration. In addition to the governance capability, the catalog gives you the ability to search and discover data assets within the enterprise.

Finally, a marketplace for data and AI solutions can foster innovation and accelerate the implementation of data-driven solutions. It allows users to discover and deploy a wide range of data and AI solutions, thereby enhancing the collaborative aspect of the strategy. The key value of the marketplace for enterprises is providing a way for them to monetize their data as data providers.

While the specific platform or technology is important, the focus should be on how these features support designing a robust and effective data collaboration framework and strategy. The goal is to address the key challenges of data collaboration and enable organizations to unlock the full potential of their data.

This section provides a framework for designing a successful data collaboration strategy in healthcare.

Goals and Scope

The first step in crafting a data collaboration strategy is to clearly define the goals and scope of the collaboration. What are the specific objectives and outcomes you hope to achieve through this collaboration? What are the use cases that this collaboration will enable? It's also important to determine the scope and boundaries of the data to be shared and co-created. This includes identifying the types of data, the volume of data, and the frequency with which the data will be shared.

For instance, a city's public health department may aim to collaborate with local hospitals to reduce the rate of hospital readmissions within 30 days of discharge. The collaboration could enable use cases such as predictive modeling of readmission risks, identification of high-risk patient groups, and evaluation of intervention effectiveness. It will also be important to identify the types of data, such as patient demographics, diagnoses, treatment plans, and readmission dates. The volume of data could range from thousands to millions of records, depending on the size of the city.

Finally, the frequency of data sharing in this case could be determined by the needs of the analysis and the capacity of both parties to manage the data. For instance, if the goal is to monitor readmission rates in real time to enable prompt intervention with high-risk patients, the data might need to be shared daily. However, if the analysis focuses on long-term trends and the effectiveness of interventions over time, then a weekly or monthly data sharing schedule might be sufficient. It's important that both the public health department and the hospitals agree on a frequency that ensures the shared data is timely and relevant, as well as manageable in terms of data privacy, storage, and processing capacities.

Partners and Roles

Once the goals and scope have been established, the subsequent phase involves the identification of potential collaborators and stakeholders. These entities should have the capacity to contribute to or derive benefits from the data collaboration. The partner identification process is crucial in any data collaboration context, including but not limited to healthcare scenarios. The collaborators could range from internal departments within your organization to external business partners—each with a unique role to play in the collaboration:

Internal departments
: Various internal departments within your organization can play significant roles and reap substantial benefits. For instance, in a healthcare context, a hospital's IT department could set up a data sharing environment with Delta Sharing and Unity Catalog to help manage the secure and efficient transfer of data between departments. This not only streamlines the process but also establishes a "single source of truth," eliminating redundancy and ensuring everyone is working with the same information.

The admissions department can track data from patient health records, including insurance and dates of service and invoicing, to help streamline the patient experience between medical departments.

The quality improvement department could also access this shared data to monitor overall hospital performance on patient care. This collaboration allows for a greater data-driven approach to improve patient care and conveys confidence to internal stakeholders.

In essence, the collaboration between internal departments, facilitated by a well-defined data collaboration strategy, can lead to a more efficient and effective organization. It ensures that all departments are aligned, working from a "single source of truth" to help make informed decisions based on the same dataset. This streamlines operations and enhances the overall efficiency and effectiveness of the organization.

External business partners

Sharing data with external partners is a routine yet crucial activity in many organizations, including those in the healthcare sector. However, it can often be a slow, cumbersome, and challenging process. The transformative power of data collaboration can revolutionize this process, making it efficient and effortless.

For instance, consider a healthcare scenario in which a hospital needs to share patient data with a research institution for a study. Traditionally, this process could be slow and complex, fraught with privacy concerns and regulatory hurdles. But with data collaboration, the hospital can share the required data securely and in real time, enhancing the efficiency of the research process.

Moreover, data collaboration also supports maintaining access audits and logs, a critical requirement for regulatory compliance in the healthcare sector. It ensures that all data access and data sharing are tracked and recorded, providing a secure and accountable system.

Similarly, consider a monetization use case in the realm of data collaboration. For instance, a healthcare provider could collaborate with a pharmaceutical company by sharing anonymized patient data on drug research and development. The

> insights derived from this data collaboration could lead to the creation of new treatments, directly benefiting the end users—the patients.
>
> This simultaneously opens a new revenue stream for the healthcare provider in which it becomes a data provider based on its unique datasets. It could charge a fee for data access and usage, creating a beneficial situation for all—the pharmaceutical company gains valuable insights for its research, the healthcare provider monetizes its data assets, and the end users potentially benefit from an improved clinical experience.
>
> This data sharing scenario illustrates how collaborating with external partners can lead to monetization opportunities, turning data into a valuable and revenue-generating asset, all while enhancing the end user experience.
>
> Data collaboration can significantly enhance data sharing across many industries, such as healthcare, by making the experience more manageable, productive, and in compliance with applicable data regulations.

In data collaboration, it's essential for partners to have a clear understanding of their respective roles. These roles can be categorized as data providers, consumers, and intermediaries. For instance, in a healthcare scenario, hospitals could be the primary data providers, while entities such as public health departments, insurance companies, and pharmaceutical companies act as data consumers. The hospital's IT department could serve as an intermediary, managing the technical aspects of data sharing. This clarity of roles helps ensure a smooth and effective data collaboration process.

However, it's important to note that these roles can vary depending on the context and scope of the collaboration. For example, with Databricks, secure collaboration can be facilitated through Delta Sharing (see Chapter 2), which provides cross-platform and cross-region sharing, and Databricks Marketplace (see Chapter 3), which provides private exchanges and internal marketplaces. Depending on the privacy-safe level required, Databricks Clean Rooms (see Chapter 4) is built on Delta Sharing and provides a secure environment for two or more parties to collaborate in a controlled way, with no direct access to their respective datasets.

Agreements and Rules

Following identification of the partners and their roles, the next crucial step in data collaboration is to negotiate and formalize the service-level agreements (SLAs) of the collaboration. This process involves addressing key issues such as data ownership, access rights, usage policies, quality standards, and compliance requirements. It's essential to create clear and transparent data sharing agreements and rules to ensure all parties have a mutual understanding and align on the terms of the collaboration. Several agreements might need to be established on the following aspects of data collaboration:

Data ownership

This refers to who owns the data and who has the right to modify or delete it, or to share it with others. In a healthcare scenario, the hospitals might retain ownership of their patient data but grant the public health department a license to use it for specific purposes.

Access rights

This pertains to who can access the data and under what conditions. For instance, a public health department might need to obtain consent from patients before accessing their personal health information. Access might also be limited to certain staff members within the hospitals or the public health department.

Usage policies

These are guidelines specifying how the data can be used. For example, in a healthcare setting, the agreement could specify that data can be used only for improving public health outcomes and not for commercial purposes.

Quality standards

These are criteria the data must meet in terms of accuracy, completeness, and timeliness. In a healthcare scenario, hospitals might need to ensure that the data they provide meets these standards.

Compliance requirements

These are legal and regulatory obligations that must be met when handling data. In a healthcare context, both parties would need to comply with privacy laws and regulations such as HIPAA in the US, which protects patients' medical records and other health information provided to health plans, doctors, hospitals, and other healthcare providers. For more information on key regulations governing data sharing, please refer back to "Key Regulations Governing Data Sharing in Different Regions and Domains" on page 16.

Platform and Tools

The choice of platform and tools can significantly impact the success of your data collaboration initiative. In addition to the critical data collaboration and sharing features that we have discussed in this book so far, including Delta Sharing, Clean Rooms, Unity Catalog, and the Marketplace, you should also consider factors such as scalability, performance, security, interoperability, ease of use, and compatibility with existing systems. The platform and tools should enable data sharing among different users and organizations, facilitating data collaboration and innovation. Unified analytics and data intelligence platforms such as Microsoft Fabric and Databricks offer a range of benefits and capabilities that enhance data collaboration. They provide scalable and cost-effective cloud services for storing and processing large volumes of data. They simplify data workflows, enhance collaboration, and unlock the full potential of data assets. They also incorporate advanced technologies such as generative AI, AI assistants, low-code/no-code solutions, and comprehensive support for machine learning workflows. Here are additional factors to consider when choosing the right data platform and tools for your collaboration strategy:

Scalability

The ability to handle large and complex datasets without compromising performance or quality is crucial. Being able to scale up or down to meet changing demands and requirements is also important. Furthermore, platforms that offer pay-as-you-go pricing models provide cost efficiency by allowing you to pay only for what you use. These are the expected capabilities of the platform and tools in use.

Openness

A recipient can be on any platform. Regardless of where your data is stored or which cloud provider you use, Delta Sharing can help you share data across platforms using connectors from pandas, Apache Spark, Power BI, Rust, and many others. You can also connect across different regions and cloud providers without any hassle. For example, you can use Delta Sharing to share data between AWS and Azure, or between the US and Europe regions. With Delta Sharing, you can collaborate with any of your global teams and partners, whether they are on Databricks or not, without worrying about access, data compatibility, or latency.

Performance

Efficient and effective data processing that delivers fast and accurate results is a key requirement. Leveraging the latest technologies and innovations, such as generative AI models for enriching data or deriving insights and recommendations, is also essential. These are the expected capabilities of the platform and tools in use.

Data security and privacy

Data security and privacy are paramount concerns that should be woven into every stage of the data lifecycle, from initial collection to storage, processing, sharing, and eventual disposal.

Data security involves protecting data from unauthorized access, modification, or loss. This includes monitoring and auditing data usage to ensure policy compliance, using tools like audit logs and usage reports. Compliance with relevant laws and regulations regarding data security is also essential. Features such as role-based access control, audit logging, and compliance certifications for standards such as HIPAA, SOC 2, and so on are expected. Data at rest can be protected using encryption services, while data in transit can be secured with SSL/TLS certificates.

Data privacy, on the other hand, focuses on ensuring that all participants are well-informed about their responsibilities regarding data privacy and that data is used in a manner that respects individual privacy rights. Modern cloud platforms offer a variety of tools that can help ensure data privacy. For instance, Databricks Delta Sharing, Databricks Clean Rooms, and Databricks Marketplace all provide a secure and scalable environment for data sharing. These tools allow for controlled access to data, ensuring that only authorized individuals can view or manipulate it.

Interoperability

Seamless integration with other systems and applications is crucial, enabling data exchange across various sources and formats. Support for multiple languages and frameworks for data processing and analysis, such as SQL, Python, R, Scala, and Java, is also essential. These capabilities are expected of the platform and tools being used.

Ease of use

Ease of use and understanding are crucial for any platform and tools, as they reduce the learning curve and boost user productivity. Features enhancing user experience, such as AI assistants for data exploration and query formulation and low-code/no-code solutions for creating data pipelines and applications, are expected. The platform should also support modern data collaboration architectures. Additionally, it should provide a user-friendly interface and tools that make it easy for users of all skill levels to collaborate on data. These include features for data exploration, visualization, and collaboration.

Compatibility

Compatibility with existing systems and applications is crucial, minimizing disruption and migration costs. Support for legacy data sources and formats is also important, as it ensures data continuity and quality. The platform should also allow connections between on-premises data sources and cloud services, supporting various data formats such as CSV, JSON, Parquet, and ORC.

Governance and compliance

Data governance and compliance encompass a broad scope that extends beyond security and privacy. It involves the efficient, effective management and use of data in a manner that is compliant with all relevant laws and regulations.

A robust data governance strategy includes mechanisms for data quality control, ensuring data consistency and reliability. This strategy involves defining data policies, rules, standards, roles, and responsibilities. Key features such as schema enforcement, transaction support, and data versioning are integral to maintaining high-quality data.

Moreover, a well-governed platform enables data monitoring, auditing, and reporting, ensuring data compliance and quality. Modern data platforms that support a unified data governance service can help you discover, catalog, classify, map, lineage, and track your data assets across different sources. These platforms enforce data policies, identify data risks, and measure data quality metrics.

Finally, compliance with relevant data protection and privacy regulations is a critical aspect of data governance. This includes features for data anonymization, consent management, and audit trails. These measures ensure that your organization's data practices are efficient, effective, and adhere to the necessary legal and regulatory standards.

Support and community

A strong support network and active community can be invaluable. The benefits include access to technical support, training resources, and a community of users who can share their experiences and insights.

Security and Governance

When you deal with sensitive or regulated data, you need to ensure there is the necessary level of data security and governance measures in place. Databricks Delta Sharing makes it easy for you to manage who can access your data and how they can use it:

Access control

Access control is a key feature of data sharing that ensures that data consumers can access only the data they are authorized to. Access control can be implemented at different levels of granularity, such as table, share, or row and column level. The latter is also known as fine-grained access control, and it allows the data provider to specify different permissions for different users or groups for each row and column of the data.

Access control can also be dynamic or static, depending on how the permissions are determined and enforced. Dynamic access control means that the permissions are evaluated at runtime, based on the data consumer's identity and other attributes. This allows the data provider to share multitenant table data with multiple data consumers, without having to create custom data slices for each one. Static access control means that the permissions are predefined and fixed and do not change based on the data consumer's identity or attributes.

Access control also enables incident mitigation, which means that the data provider can quickly revoke or modify the permissions of any data consumer in case of a breach or a change in the data sharing agreement. Thus the data provider can protect the data from unauthorized or malicious access.

Network security and encryption

When you share data with other parties, you need to ensure that the data is protected from unauthorized or malicious access, both in transit and at rest. Databricks Delta Sharing provides several features and best practices to help you secure your data sharing network, such as:

End-to-end encryption

Delta Sharing uses TLS to encrypt the data in transit between the data provider, the Delta Sharing server, and the data recipient. Delta Sharing also supports encryption at rest for the data stored in the data lake, using the native encryption features of the cloud storage providers. You can also encrypt your data using industry-standard protocols and algorithms, such as AES-256 for data storage and HTTPS and TLS for data transmission. For example, if you are a healthcare provider who wants to share anonymized patient data with a research institute for a medical study, you can create a share of your patient data and grant read-only access to the research institute. You can also revoke access to the share at any time, if needed.

Short-lived credentials

Delta Sharing uses short-lived credentials, such as presigned URLs, to grant temporary access to the data. These credentials expire after a specified time, which the data provider can configure. This reduces the risk of credential leakage or misuse by unauthorized parties.

IP access lists

Delta Sharing allows the data provider to configure IP access lists, which are rules that specify which IP addresses are allowed or denied access to the shared data. This helps to restrict access to the data to only trusted parties and block any unwanted or suspicious requests.

Network restrictions

Delta Sharing also allows the data provider to configure network restrictions on the storage account in which the data is stored, such as firewall rules or virtual network settings. Such restrictions help isolate the data from the public internet and limit access to the data to only authorized networks or devices.

Audit and compliance

You can use logs and reports to monitor and audit your data sharing activities. You can also comply with various data regulations and standards, such as HIPAA, GDPR, and ISO 27001, by using features such as data anonymization, data retention, and data deletion. For example, you can use logs and reports to track and audit who accessed your patient data, when they accessed it, and what they did with it. You can also generate reports on data usage and performance. In addition, you can comply with HIPAA regulations by anonymizing your patient data, retaining your data for a specified period, and deleting your data when it is no longer needed.

Unity Catalog and security best practices

Use Unity Catalog to manage and organize your data sharing resources by providing a unified view of all your shares, such as tables, AI models, and volumes, regardless of where they are stored or which cloud provider you use. You can also use Unity Catalog to implement security best practices for data sharing by providing features such as authentication, authorization, encryption, and auditing. For example, you can use Unity Catalog to authenticate the research institute, authorize its access to the share, encrypt the metadata and data, and audit the data sharing activities. You can also view and manage all your shares and tables using Unity Catalog.

Databricks Data Intelligence Platform for Data Collaboration

Databricks developed the Data Intelligence Platform to allow your entire organization to use data and AI. It is built on a data lakehouse to provide an open, unified foundation for all data and governance and is powered by a Data Intelligence Engine. Built on open source and open standards, a lakehouse simplifies your data estate by eliminating the silos that historically complicate data and AI. These platforms enable data collaboration across different teams within an organization by providing such features as:

- Automated management and optimization, which improves data quality, performance, and efficiency. For example, a data provider can set up data quality rules and alerts, such as checking for missing values, duplicates, outliers, and anomalies, and be notified when the data quality changes or degrades. A data recipient can enable auto-refresh for their Delta Shares and automatically get the latest data updates without manually reloading the data. Additionally, both data providers and recipients can benefit from automatic indexing and query optimization, which automatically create and maintain indexes for the data, tune the query parameters, and optimize the query execution plan, reducing the query latency and cost.
- Enhanced governance and privacy, which protect and control data access and usage. For example, a data provider can define data policies and permissions (such as who can view, edit, or share data), and apply data masking, encryption, or anonymization techniques to protect sensitive data. A data recipient can comply with the data policies and permissions and respect the data privacy and security of the data provider. Moreover, both data providers and recipients can use data classification to label and filter data based on its level of sensitivity, and use data governance tools to set up data access rules and roles, audit the data and AI activities, track the data and AI lineage, and discover and catalog the data and AI assets.

- First-class support for data and AI workloads, which allows users to leverage data insights and AI capabilities in their applications. For example, a data provider can use data and AI services such as data engineering, data science, machine learning, and analytics to build, train, test, and deploy data and AI applications, including recommender systems, fraud detection, sentiment analysis, and more. A data recipient can use data and AI services, such as data integration, data visualization, machine learning, and analytics, to consume, analyze, and enrich the data and AI applications, such as dashboards, reports, insights, and predictions.

The Databricks Data Intelligence Platform is powered by the lakehouse, a unified system that combines the best of data lakes and data warehouses. The lakehouse allows users to store and manage all types of data across the enterprise, from structured to unstructured, from batch to streaming, from historical to real-time. The lakehouse was pioneered by Databricks, offering a unique platform with a unified governance layer and a single query engine.[1] This modern unified data and analytics platform offers scalable and cost-effective solutions for storing and processing large volumes of data, including pay-as-you-go models that eliminate the need for large upfront infrastructure investment. Moreover, the lakehouse decouples storage from compute, allowing users to scale their resources independently based on their specific needs, leading to more efficient data management and cost savings.

Designing a Data Collaboration Architecture

When it comes to designing a data collaboration architecture, technical and organizational aspects are paramount. The architecture needs to cater to the technical necessities of data ingestion, transformation, analysis, visualization, and dissemination. It should resonate simultaneously with the organization's objectives, culture, and capabilities. A solid data collaboration architecture should:

[1] Michael Armbrust, Ali Ghodsi, Reynold Xin, and Matei Zaharia, "Lakehouse: A New Generation of Open Platforms That Unify Data Warehousing and Advanced Analytics," in *11th Annual Conference on Innovative Data Systems Research (CIDR '21)*, January 11–15, 2021, *https://oreil.ly/YFniN*.

- Be open and flexible, allowing for interoperability and integration of various data types, formats, and sources
- Be platform-agnostic, enabling data sharing and access across different cloud providers and regions, without compromising data quality or security
- Support live data sharing, avoiding the need for data replication or transformation, which can introduce latency and inconsistency
- Align with the organization's objectives, culture, and capabilities, fostering a data-driven mindset and culture

One emerging data collaboration architecture is a data mesh, an organizational framework that decentralizes data ownership and management, empowering domain teams to own and manage the data they generate. A data mesh handles data as a product rather than a byproduct and enables data sharing and collaboration through standardized protocols and interfaces. A data mesh reduces the burden on centralized data teams and increases the agility and autonomy of domain teams, paving the way for more efficient and effective data collaboration.

Data Mesh

A data mesh is a distributed data architecture that organizes data around business domains rather than around technical functions. Each domain takes ownership and manages its own data, presenting it as a data product via standardized APIs and protocols. These data products are self-describing, discoverable, and interoperable, facilitating easy access and usage for data consumers. A data mesh also ensures data governance and quality at the source, making the data reliable and compliant.

The data mesh model enhances data sharing and collaboration by enabling domain teams to function as both data providers and consumers, eliminating the need for a centralized data platform and team. This model gives data providers the autonomy and motivation to offer high-quality, relevant, and timely data to their consumers, who can access and use the data in a self-service manner.

Data mesh integrates with modern data sharing technologies, such as Databricks Marketplace, Delta Sharing, and Databricks Clean

Rooms. These tools, along with microservices[2] and generative AI, facilitate data sharing and collaboration at scale, both internally and externally.

Adopting a data mesh allows organizations to establish a decentralized, domain-oriented, product-based, and self-governing data ecosystem. This aids in achieving data sharing and collaboration objectives, such as enhancing data availability, accessibility, usability, and value. Furthermore, data mesh can help organizations cultivate a data-driven culture in which data is viewed as a strategic asset and a catalyst for innovation. Figure 5-4 illustrates the data mesh operating model.

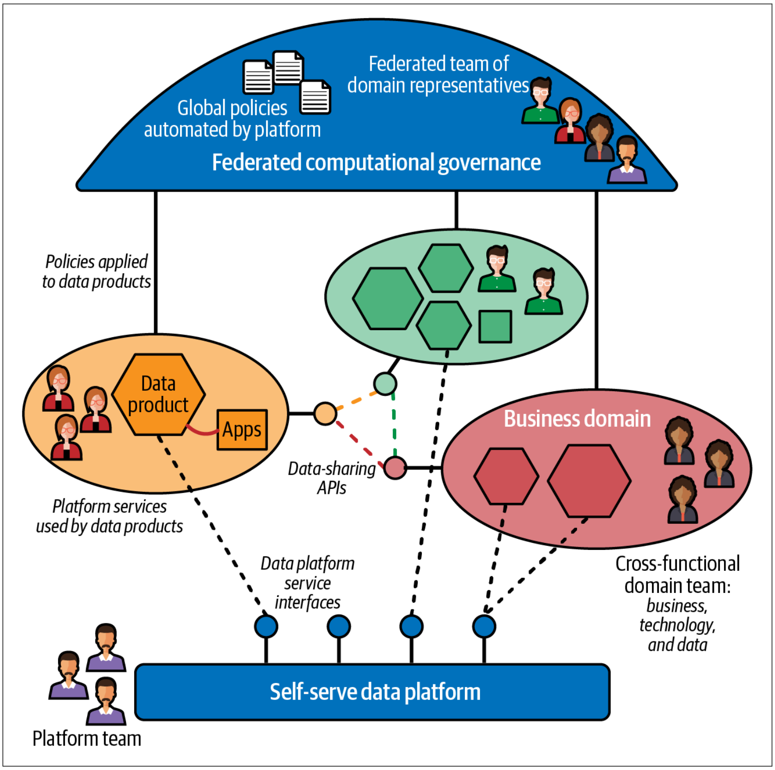


Figure 5-4. Data mesh operating model

[2] Microservices is an architectural style in software development that structures an application as a collection of small, autonomous services, each running its own process and communicating with lightweight mechanisms, often over APIs.

Adapting Architectures and Continuous Improvement

An effective data collaboration architecture is not a one-size-fits-all solution but rather a dynamic and evolving system that adapts to the changing needs and context of the organization. Therefore, it is important to continuously monitor and improve the data collaboration architecture through the use of metrics, feedback, and best practices. The cornerstone of a successful data collaboration and sharing architecture is tailoring it to your organization's unique needs.

The focus should be on fostering data collaboration and sharing—whether internally, externally, or both, depending on where an organization is in its data sharing and collaboration journey (see Figure 5-5), which is critical for leveraging collective intelligence to drive innovation and secure a competitive advantage.

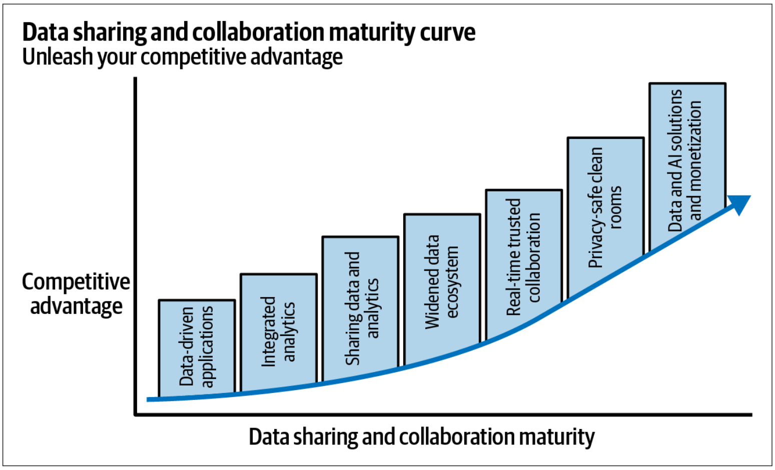


Figure 5-5. Data sharing and collaboration maturity curve

In addition to the data architecture, it's crucial to consider the broader technical infrastructure. Employing a version control system like Git to track modifications in your code, models, and configurations will not only facilitate collaboration and reproducibility but also ensure safe data sharing and allow for the rollback of changes.

You should incorporate DevOps practices and continuous integration/continuous deployment (CI/CD) pipelines to automate the testing, integration, deployment, and monitoring of your data pipelines and models. This will enhance your data collaboration efforts by ensuring data integrity and reliability. A tool such as Azure DevOps, GitHub Actions, or Jenkins can be instrumental in this process.

Remember, designing a data collaboration and sharing architecture is not a one-off task but a continuous journey. The architecture should evolve with your changing data needs, technology landscape, and organizational context. This adaptability is key to maintaining effective data collaboration and sharing over time.

Scope and Scale of Implementation

Data collaboration is a key enabler for data-driven innovation, allowing organizations to leverage data from diverse sources and partners to generate insights and value. However, data collaboration also poses significant challenges, such as data security, governance, compatibility, and performance. Organizations need a robust and flexible data sharing strategy that can support their data collaboration goals to address these challenges. Delta Sharing provides such a strategy, enabling secure and efficient data sharing across platforms, clouds, and regions. In this section, we will discuss implementing Delta Sharing for data collaboration, using a phased approach involving a technical pilot and a full-scale rollout. We will also highlight the ease of implementation and the benefits of Delta Sharing for data collaboration.

Ease of Implementation

Delta Sharing is designed for ease of implementation, integrating seamlessly with existing data infrastructure and requiring minimal changes to current workflows. It provides a simple REST API that facilitates seamless data collaboration by allowing data providers to share their data and recipients to read the shared data. The platform supports AI and machine learning workflows, enabling direct access to shared data for advanced data analysis. Furthermore, Delta Sharing ensures data freshness by allowing real-time data sharing, making it particularly useful for use cases that require timely analytics. This comprehensive and scalable solution reduces the time and effort required to adopt data collaboration, making it an attractive tool for both data providers and recipients.

Phased Implementation Approach

Implementing Delta Sharing for data collaboration can be effectively achieved through a phased approach. Initially, a technical pilot is conducted to test the strategy in a controlled environment, assessing the technical feasibility and operational impact of new tools and practices. This pilot involves using key features of the Data Intelligence Platform such as Delta Sharing, Unity Catalog, Data Marketplace, and Databricks Clean Rooms with a representative group of target users. Clear objectives and expectations are set for the pilot, and its progress and outcomes are closely monitored. Feedback from the pilot users is gathered and used to refine the strategy, with adjustments made based on the feedback and pilot results. This process helps identify potential challenges and mitigate risks and ensures a smoother transition when the data collaboration strategy is rolled out across the organization, maximizing its value.

Implementing a technical pilot for Delta Sharing involves several steps from both the provider's and the recipient's perspectives. Here's a detailed approach:

Provider perspective

1. *Identify data for sharing.*
   Identify the datasets that will be shared during the pilot. These should be representative of the data that will be shared in the full-scale implementation.

2. *Set up Delta Sharing.*
   Install and configure Delta Sharing on your data platform. This step involves setting up the Delta Sharing server and configuring the necessary permissions and security settings.

3. *Prepare data for sharing.*
   Preparing the identified datasets for sharing involves converting the data into Delta Lake format (if it is not already in this format), partitioning the data for efficient sharing, and setting up the necessary sharing profiles.

4. *Test data sharing.*
   Share the prepared datasets with the pilot users (recipients) and ensure they can access and use the data as expected.

Recipient perspective

1. *Set up the Delta Sharing client.*
   Install and configure the Delta Sharing client to access the shared data. This step involves setting up the necessary permissions and security settings on the client side.

2. *Access and integrate shared data.*
   Access the shared data using the Delta Sharing client. Test various operations such as reading the data, querying the data, and integrating the data to enrich other datasets.

3. *Provide feedback.*
   Provide feedback on the data sharing process, the usability of the shared data, and any issues encountered. This feedback will be used to refine and improve the data sharing strategy.

After the pilot, both providers and recipients should review its outcomes, make necessary adjustments to the data sharing strategy, and plan for the full-scale implementation. This phased approach helps ensure the successful implementation of Delta Sharing for data collaboration.

Full-Scale Rollout

After the technical pilot has successfully been completed, the data sharing strategy can be scaled up to the full organization level. This involves expanding the scope and scale of the data sharing, both in terms of the number and variety of datasets shared and the number and diversity of users accessing the shared data. The full-scale rollout also involves ensuring the sustainability and governance of the data sharing, as well as measuring and communicating the value and impact of the data collaboration. Here are some key steps for the full-scale rollout:

For data providers

1. *Expand data sharing.*
   Identify and prepare additional datasets for sharing, based on the feedback and results from the pilot. This may involve converting more data into Delta Lake format, partitioning and optimizing the data for sharing, and setting up more sharing profiles. Share the data with more users across the organization, as well as with external partners and clients if applicable.

2. *Ensure data freshness and security (ongoing).*
   Ensure that the shared data is being updated and reflects the latest version of the data. This may involve validating the data, as well as applying data quality rules and metrics. Also ensure that the shared data is secure and compliant with the relevant policies and regulations. This may involve encrypting, masking, or anonymizing the data, as well as applying data access controls and auditing mechanisms.

3. *Monitor and optimize data sharing (ongoing).*
   Monitor the performance and usage of the shared data and identify any issues or bottlenecks that may affect the data sharing. This may involve tracking metrics such as data availability, latency, throughput, egress costs, and user feedback and satisfaction. Optimize the data sharing process by making necessary adjustments and improvements, such as tuning the configurations or applying best practices.

For data recipients

1. *Access shared data.*
   Access the shared data using the Delta Sharing client. Test various operations such as reading, querying, and integrating the data with other datasets.

2. *Provide feedback (ongoing).*
   Provide feedback on the data sharing process, the usability of the shared data, and any issues encountered. This feedback will be used to refine and improve the data sharing strategy.

3. *Measure and communicate value (ongoing).*
   Measure the value and impact of the data sharing and collaboration, in terms of both technical and business outcomes. This may involve defining and tracking key performance indicators (KPIs) such as data quality, data usage, data insights, data-driven decisions, and data-driven actions. Communicate the value and impact of the data sharing and collaboration to the internal and external stakeholders. This may involve creating and sharing reports, dashboards, stories, or testimonials that showcase the benefits and achievements of data sharing and collaboration.

By following this phased approach, both providers and recipients can ensure a successful full-scale rollout of Delta Sharing for data collaboration.

Success Metrics and Goals

To gauge the effectiveness of your data collaboration strategy, it's essential to establish and monitor KPIs that align with your organization's specific objectives. These KPIs should reflect the success metrics of your business. For instance, an airline might focus on fuel consumption and flight plan optimization as KPIs for data sharing.

While metrics such as the number of active collaborations, the volume of data shared, and understanding derived from shared data can provide valuable insights into the engagement and volume of data collaboration, they should be contextualized for your business goals.

Similarly, data quality metrics such as accuracy, completeness, consistency, timeliness, and relevance should be assessed in relation to how they contribute to achieving your business objectives.

User engagement metrics, such as the number of active users, frequency of use, types of actions performed, and user feedback, can offer insights into how users interact with data collaboration tools. However, these should also be tied back to how they support your business goals.

Ultimately, the primary benefit of data collaboration is to drive positive business outcomes. Therefore, evaluating business outcomes such as revenue growth, cost savings, customer satisfaction, and operational efficiency is crucial. These metrics should directly reflect the impact of data collaboration on your organization's bottom line:

Revenue growth

Measures how much data collaboration contributes to increasing your organization's income and profits by creating new products, services, or markets or improving existing ones.

Cost savings

Indicates how much data collaboration helps to reduce your organization's expenses and inefficiencies by optimizing processes, resources, or operations.

Customer satisfaction

Reflects how much data collaboration improves your organization's relationship with your customers by meeting their needs, expectations, and preferences or enhancing their experience or loyalty.

Operational efficiency

Shows how much data collaboration enhances your organization's performance and productivity by streamlining workflows, improving quality, or reducing errors or risks.

Last, data collaboration often results in learning and growth opportunities. Metrics such as new skills acquired by staff, new partnerships formed, and new areas of research explored can be valuable indicators of the broader benefits of data collaboration. However, these should also be considered in the context of how they contribute to your organization's strategic objectives.

Remember, the key is to use these metrics to define your own KPIs based on your specific business goals. This ensures that your data collaboration strategy is robust, efficient, and aligned with your organization's vision and mission.

Change Management in Data Collaboration

Embarking on a data collaboration strategy can usher in transformative changes within an organization. These changes can span from the adoption of innovative technologies to the modification of existing workflows and processes.

Data collaboration with Databricks Delta Sharing requires effective change management to ensure a smooth transition and optimal outcomes. Change management involves a structured and systematic approach to transitioning individuals, teams, and organizations from their current state to a desired future state. This approach encompasses a range of activities, such as assessing the need for change, developing a change management plan, communicating and engaging with stakeholders, implementing the change, and evaluating its impact. These activities can be supported by leveraging features of modern data intelligence platforms such as Delta Sharing, Unity Catalog, Databricks Marketplace, and Databricks Clean Rooms.

Securing stakeholder buy-in is critical in managing change when implementing a data collaboration strategy. This process involves presenting the features along with the needs they fill to top management, as well as to those who will be directly affected by the changes.

From the provider's perspective, an efficient data collaboration strategy enables secure and efficient data sharing, centralized management and organization of data assets, wider reach and impact of data, and privacy-safe sharing of sensitive data. For example, using Delta Sharing, a provider can share real-time data from their data lake with multiple recipients, without the need for data movement or duplication. This reduces the complexity and cost of data sharing and ensures data quality and security. Using Unity Catalog, a provider can manage and organize their data assets in a single location and categorize them based on their content. This makes it easier for recipients to find and access the data they need. Using Databricks Marketplace, a provider can publish their datasets for others to discover and use and can increase the value and visibility of their data. Using Databricks Clean Rooms, a provider can share sensitive data in a secure environment in which recipients can access the data without compromising privacy or compliance.

From the recipient's perspective, data collaboration enables real-time access to shared data, easy discovery and integration of data, access to new and valuable datasets, and secure processing of sensitive data. For example, using Delta Sharing, a recipient can access shared data in real time and work with the most current and accurate data. Using Unity Catalog, a recipient can easily discover and access shared datasets and integrate them into their existing workflows. Using Databricks Marketplace, a recipient can discover new datasets that could enhance their work and access them through a simple interface. Using Databricks Clean Rooms, a recipient can work with sensitive data in a secure environment without compromising privacy or compliance.

The "any cloud, any region" capability of Delta Sharing is another compelling benefit. This feature allows data to be shared and accessed across different cloud environments and geographical regions, providing flexibility, scalability, and resilience. It also supports data sovereignty and compliance requirements, which can be crucial for stakeholders.

You could also discuss the potential for monetization through the data marketplace, which could be a compelling benefit for stakeholders. Additionally, explaining how the platform supports both data providers and consumers can help stakeholders understand its value. Data sharing and collaboration using Databricks Delta Sharing requires clear communication about the benefits and challenges of this approach. Stakeholders may have concerns or objections around data security, governance, or performance. Providing reassurances and addressing these issues can help secure their buy-in. For example, Databricks Delta Sharing permits data providers to control what data is shared and with whom, and it allows data recipients to access shared data in real time and across platforms. It's also important to align the data sharing and collaboration strategy with the organization's vision and goals and demonstrate how it will improve the quality and efficiency of data-related work. For example, Databricks Delta Sharing can help organizations achieve cost savings, revenue growth, and customer satisfaction by enabling data-driven innovation and insights.

By focusing on these areas, you can strengthen stakeholder buy-in and pave the way for the successful implementation of your data collaboration strategy. One of the remarkable aspects of data sharing is its transparency to end users. They can consume external data without even realizing it, as everything operates as if the data were local. This seamless integration reduces the need for extensive training across various personas. In fact, it may be sufficient to train only the data administrators, further simplifying the process and enhancing the efficiency of collaboration.

From a provider's perspective, technical change management is also a critical aspect of implementing a data collaboration strategy, involving the use of tools and practices such as CI/CD pipelines, version control, and DevOps to automate and streamline the data collaboration workflows. From a recipient's perspective, the principles of CI/CD apply seamlessly to shared data as if it were local. This uniformity in operation, irrespective of the data's origin, ensures a consistent and efficient workflow. It further underscores the advantage of eliminating geographical and organizational boundaries in data collaboration.

Finally, don't overlook learning and growth opportunities that often result from data collaboration. Data providers have opportunities to learn how to effectively share and manage their data assets. Delta

Sharing enhances providers' understanding of data sovereignty and compliance. Unity Catalog offers a unified view of all data assets, enabling providers to better organize and manage their data. Databricks Marketplace equips providers with insights into the value of their data and potential monetization opportunities.

For data consumers, these tools offer opportunities to learn how to discover, access, and analyze data more effectively. Delta Sharing enables consumers to access shared data in real time, enhancing their ability to perform timely analytics. Unity Catalog makes it easier for consumers to discover and access the data they need, improving their efficiency and productivity. Databricks Marketplace offers a wide range of datasets for purchase, expanding consumers' data sources and analytical possibilities.

Databricks Clean Rooms offers a secure environment for joint data analysis without exposing sensitive data, providing providers and consumers a safe space to collaborate and learn from each other.

Remember, learning and growth are not one-time events but ongoing processes. By embracing these opportunities, providers and consumers can ensure they are well-equipped to navigate the changes brought about by the data collaboration strategy and maximize its value.

Summary

Crafting an effective data collaboration strategy is a multifaceted process. It requires a deep understanding of the technological landscape, the ability to navigate challenges, and the application of data collaboration best practices. The Databricks Data Intelligence Platform can help organizations create secure, efficient, and scalable data collaborations.

Databricks provides an open and secure platform for data collaboration. Delta Sharing provides an open source approach to data sharing across clouds, platforms, and regions. Delta Sharing powers Databricks Marketplace, which opens up new opportunities for innovation and monetization, as well as Databricks Clean Rooms, a privacy-safe collaboration environment for customers and partners. On the Data Intelligence Platform, all of this is secured and governed by Unity Catalog, which provides organizations with a seamless governance layer and promotes data discoverability and secure access.

These features streamline data collaboration workflows and offer significant benefits for various stakeholders. Data providers can manage and monetize their data assets more effectively, data consumers can access a wider range of data sources and analytical tools, and regulatory entities can ensure data compliance more efficiently.

However, technology is just one piece of the puzzle. Organizations must also consider how to measure success, manage change, and continuously improve their strategies to stay competitive in the fast-paced world of big data.

This chapter equips you with the knowledge and tools to begin crafting a data sharing and collaboration strategy for your organization. Remember, the ultimate goal is not just to share data but to use this shared data to drive innovation, create value, and achieve your business objectives.

The future of data collaboration is being shaped by several emerging trends, including AI and ML for data analysis, blockchain for secure data sharing, real-time data sharing, data marketplaces, and privacy-preserving data collaboration. Generative AI is also becoming more prevalent, opening up new possibilities for data analysis and insight generation. These advancements are creating new opportunities and challenges for organizations looking to leverage the power of data.

The potential impact of a well-crafted data sharing and collaboration strategy on business performance and innovation is immense. By securely sharing and collaborating on data using modern data and advanced analytics cloud tools and platforms, businesses can unlock new insights and capabilities, improve their decision making, and drive innovation. As technology evolves, businesses will have even more options for securely sharing and collaborating on data. By staying up to date with these developments and continuously refining their data sharing and collaboration strategies, businesses can continue to achieve their goals while maintaining the highest data protection standards.

CHAPTER 6
Empowering Data Sharing Excellence

Throughout this book, you have learned that Delta Sharing offers a secure and controlled environment for sharing data across departments and with external partners, eliminating the need for data replication or movement. It enhances collaboration and provides deeper insights into data. Key features include Clean Rooms, which offers secure environments for sharing sensitive data in compliance with privacy requirements. The Marketplace serves as a hub for data products, facilitating easy access and eliminating complex procurement processes. Data catalogs act as a centralized library for shared data assets, improving data governance and security. Data quality checks ensure the validity and accuracy of shared data. Data lineage provides transparency in data sharing practices by tracking the origin and transformation of data. Data notifications keep users informed about changes in shared data. Data APIs allow for easy integration of shared data with other applications. By leveraging these features within the Lakehouse data platform, organizations can turn data into actionable insights in a secure, scalable, and real-time data sharing environment.

Excellence in data sharing can be defined as the ability to share data in a way that maximizes its value while minimizing risks. This includes ensuring the quality and accuracy of shared data, protecting sensitive information, complying with relevant regulations, and enabling effective collaboration among data users. Excellence is not just about having advanced technologies; it's about using these technologies to drive meaningful outcomes.

In today's data-driven world, achieving excellence in data sharing can provide organizations with a significant competitive advantage. It can enable them to uncover valuable insights, make informed decisions, innovate faster, and deliver superior customer experiences. In this final chapter, you'll learn about the key components for data sharing excellence and how to effectively overcome challenges to achieve this excellence.

Key Components for Data Sharing Excellence

In today's fast-paced business environment, achieving data sharing excellence is not just a luxury but a necessity. With the rise of big data and advanced analytics, organizations that can effectively share and analyze their data are better positioned to identify trends, anticipate customer needs, and respond quickly to changing market conditions. There are many examples of successful data sharing initiatives that have had a significant impact. For instance, the Bloomberg American Health Initiative has published a guide on successful data partnerships,[1] and MIT Sloan has made a case for building a data sharing culture in companies.[2] These examples highlight the importance of strategic partnerships, sustainability, links to action, and effective data communication.

To achieve data sharing excellence, organizations need to focus on five key components—holistic data stewardship, constructive teamwork approaches, optimal utilization of Delta Sharing, data protection and confidentiality, and data integrity management:

Holistic data stewardship

This is the process of setting up guidelines and protocols to safeguard and administer data. An example of this in practice is Spotify's adherence to privacy regulations, the use of customer insights by its product management teams, and its optimization of data quality through control mechanisms.

[1] Amanda Latimore and Sara Whaley, "A Quick Guide to Successful Data Partnerships," Johns Hopkins Bloomberg American Health Initiative, February 24, 2021, *https://oreil.ly/wUrJm*.

[2] Brian Eastwood, "The Case for Building a Data-Sharing Culture in Your Company," MIT Sloan School of Management, September 9, 2021, *https://oreil.ly/YHW1p*.

Constructive teamwork approaches

These approaches foster an environment of openness and trust, encouraging interdisciplinary collaboration. For instance, companies can collaborate with competitors within their industry to attain shared objectives such as gaining deeper customer understanding or identifying fraud trends across the sector. Another approach is behavior modeling, in which leaders demonstrate to employees how to collaborate effectively.

Optimal utilization of Delta Sharing

Delta Sharing enables secure and efficient data exchange. For example, a financial data provider was able to reduce operational inefficiencies in its traditional data delivery channels and provide end customers with seamless access to extensive new datasets using Delta Sharing. Similarly, a major retailer was able to share product data with partners effortlessly, despite not being on the same data sharing or cloud computing platform.

Data protection and confidentiality

It's crucial to implement strong security protocols and respect privacy rights. Methods such as data anonymization based on generalization, data encryption based on cryptography, data disturbance based on noise, or a combination of those techniques are typically used for privacy protection. Emerging privacy-preserving technologies, such as fully homomorphic encryption (FHE) and differential privacy, allow for the sharing of encrypted data and computations on that data without the need for decryption.

Data integrity management

This involves ensuring the precision, completeness, uniformity, and dependability of data. It's vital because substandard data quality can result in incorrect insights, misguided strategies, and potential regulatory compliance issues. Data quality tools, for example, can be used to validate, cleanse, transform, and manage data, ensuring it's suitable for its intended uses.

Overcoming Challenges to Achieve Excellence

Achieving excellence in data sharing is indeed a multifaceted process that necessitates strategic planning and execution. It is a journey fraught with challenges, including the existence of data silos, the need to establish trust, and the imperative to ensure security and privacy. These hurdles can seem daunting, but they are not insurmountable. By addressing these challenges head-on, organizations can pave the way toward data sharing excellence. The following six key steps provide a roadmap for this journey, offering practical solutions to overcome these obstacles and unlock the full potential of data sharing. Let's delve into these steps and explore how they can transform the data sharing landscape in your organization:

1. Dismantle data silos that prevent the free flow of information. Promote a culture of sharing data and utilizing technologies that enable secure data access across various teams.
2. Establish trust by implementing clear data usage policies and maintaining transparency in all operations. Delta Sharing can serve as an effective tool in this regard, facilitating seamless and secure data sharing.
3. Ensure security and privacy through robust measures such as encryption and access controls, along with technologies designed to protect sensitive information.
4. Maintain data quality by implementing checks at every stage of the data lifecycle and using tools that automate data cleaning processes.
5. Manage the complexity of data sharing by investing in the right tools and training for teams. An open and transparent organizational culture can further support data sharing initiatives.
6. Stay abreast of regulatory developments and have a compliance program in place to ensure that data sharing practices align with legal requirements.

Examples of Data Sharing Excellence

Data sharing has enabled several industries to tap into the power of big data and gain actionable insights. Uber, for example, has used data sharing to match users with the nearest driver in just seconds, enhancing its service delivery and giving it a competitive edge in the market.

In the retail sector, stores operating both online and in physical locations have a lot of data to deal with. Their secret to tracking performance and making informed decisions is in centralizing this data, no matter which store or employee entered it. This integration allows these stores to manage crucial metrics like inventory, labor hours, and sales across all their channels and outlets.

Healthcare is another area in which data sharing plays a critical role. Comprehensive patient care requires as much information as possible. When data is spread across different systems, it compromises the quality of care. But when patient data is integrated into a comprehensive record, healthcare can be transformed by controlling costs, improving outcomes, and promoting overall wellness.

Fraud is a significant challenge in finance. Banks and other financial institutions can identify, eliminate, and prevent instances of fraud if all their data is integrated. Once the data is integrated, AI can mine it for anomalies and outliers, often catching fraudulent activities before they affect the customer.

Even federal agencies have seen the benefits of data sharing. The US government's aid efforts in response to Hurricanes Irma and Maria in Puerto Rico were hampered by imperfect address data for the island. In the aftermath, emergency responders gathered to enhance the utility of Puerto Rico address data and share best practices for using what information is currently available. The Department of Energy's National Nuclear Security Administration (NNSA) adopted a data-driven, risk-informed strategy to better assess risks, prioritize investments, and cost-effectively modernize its aging nuclear infrastructure. The Federal CDO Council worked with several US departments to develop a Diversity Profile Dashboard and explore the value of shared HR decision support across agencies.

These examples highlight how operational excellence can be achieved through secure and efficient data sharing practices, and they emphasize the need for organizations to invest in the right tools and technologies that facilitate such practices.

The Future of Data Sharing

Driven by innovations such as Delta Sharing, which is paving the way for more secure, efficient, and collaborative data sharing, the future of data sharing is one in which organizations can share data seamlessly and securely, unlocking the full potential of their data. Effective data sharing can enable organizations to drive innovation, improve collaboration, and make a positive impact on society.

According to a recent survey by Forrester Research, more than 70% of global data and analytics decision makers are expanding their ability to use external data, and another 17% plan to do so within the next 12 months.[3] Gartner predicts that by 2023, organizations that promote data sharing will outperform their peers in most business metrics.[4] Data sharing can increase efficiency and lower costs, as well as broaden research collaboration and secure intellectual property. These opportunities can foster collaborative research efforts to achieve a common goal, such as bringing a lifesaving innovation to market more quickly. Data sharing also brings the opportunity for innovative and growth-focused organizations to monetize data, AI, ML, and apps that can be delivered to customers securely and instantly.

Cloud-based data sharing platforms have ushered in a new era of effortless data exchange for organizations. These innovative data marketplaces operate under a data-sharing-as-a-service model, empowering subscribers to efficiently manage, customize, and monetize their data offerings. Utilizing platform-ensured clean rooms, organizations can securely amalgamate their data resources, fostering collaborative analysis. This ecosystem enables subscribers to aggregate data and provide data access to fellow participants, offering tailored insights into diverse market segments, products, or research endeavors.

The landscape of data sharing is rapidly evolving, driven by a surge in use cases as organizations seize the opportunity to maximize the value of their data assets. In the nascent phase of the data marketplace sector, both startups and cloud providers are engineering inventive solutions that unlock promising avenues. Amid escalating demand for external data, and propelled by trends like data democratization, growth, and digital transformation, data is transforming into a pivotal business asset, ripe for trading, sharing, and strategic collaboration. The platform that seamlessly facilitates this exchange stands poised to potentially set an industry-wide benchmark, transcending individual data verticals to reshape entire markets through a paradigm-shifting approach to data sharing.

[3] Jennifer Belissent, "CDOs Wanted: Dedicated, Expanded Data Insights Leadership," Forrester (blog), January 8, 2021, *https://oreil.ly/Xo0YB*.

[4] Laurence Goasduff, "Data Sharing Is a Business Necessity to Accelerate Digital Business," Gartner, May 20, 2021, *https://oreil.ly/tW2pX*.

Additionally, Delta Sharing is constantly evolving to support more features and capabilities that can enhance your data sharing experience. Some of the future developments in Delta Sharing include:

Support for more data formats
: Delta Sharing will support more data formats beyond Delta Lake, such as Parquet, CSV, and JSON. This will enable more flexibility and compatibility for data sharing across different platforms and tools.

Support for more cloud providers
: Delta Sharing will support more cloud providers beyond AWS, such as Azure and GCP. This will enable more scalability and availability for data sharing across different regions and environments.

Support for more authentication methods
: Delta Sharing will support more authentication methods beyond OAuth 2.0, such as SAML and Kerberos. This will enable more security and convenience for data sharing across different organizations and users.

Call to Action: Your Path to Excellence

Data sharing and collaboration are essential tools for unlocking the full potential of data-driven innovation. With advanced technologies like Delta Sharing, organizations can share data securely and efficiently, opening new opportunities and creating unprecedented value. Whether it means gaining deeper customer insights, detecting fraud patterns, or developing lifesaving solutions, data sharing enables organizations to achieve common goals and even monetize their data assets.

We hope this book not only has informed you about the principles, practices, and tools that lead to data sharing excellence but also has inspired you to act. The real-world examples presented in this book demonstrate the tangible benefits of achieving excellence in data sharing.

But inspiration and knowledge are just the beginning. The next step is action. It is time to apply what you have learned to your own organization. Start by identifying potential areas for data collaboration in your organization or opportunities for using Delta Sharing. Consider conducting a pilot project to test the waters. Seek feedback from all stakeholders and be ready to learn and adjust your strategy as you go.

Remember, the path to data sharing excellence is not a sprint but a marathon, requiring ongoing effort and adaptation. But with every step on this path, you will be unlocking the full potential of your data, driving innovation, and creating value for your organization.

About the Author

Ron L'Esteve is a professional in the technology sector, with a focus on the cloud, data, and AI tech industries. He contributes to the development of capabilities within hybrid and multi-cloud platforms. As an early adopter, Ron is involved in the field of emerging cloud tech services. His practical approach to leadership has helped in the discovery of digital transformation opportunities and their progression from idea to implementation. In addition to his work in technology, Ron has written three books published by Apress. He also shares his technical knowledge through MSSQLTips.com, where he has published over 100 articles.